Establishing Physics Concept Inventories Using Free-Response Questions

Thesis

Establishing Physics Concept Inventories Using Free-Response Questions

Mark Alfred Joseph Parker

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Physics Discipline

School of Physical Sciences

Faculty of STEM

The Open University, UK

28th July 2020

Abstract

Concepts are important building blocks for understanding complicated topics and entire disciplines such as physics. The idea of an inventory of concepts has been proposed as the basis for investigating the readiness of students to develop their understanding. Hestenes et al. introduced concept inventories in physics using the multiple-choice question format. There is interest in using a less-constrained free-response format with computer-automated marking to enable more efficient use of concept inventories. The adaptation of Hestenes' Force Concept Inventory (FCI) for use with free-response, computer-marked format is the subject of this thesis. This study establishes the Alternative Mechanics Survey (AMS), a free-response computer-marked version of the FCI and validates it for use as an alternative to the multiple-choice FCI.

The AMS was subject to validity testing with a pilot group through usability tests followed by semi-structured interviews, which were analyzed using Thematic Analysis. This established that the AMS structure involving the free-response format was viable. Classical Test Theory (CTT) was used to test for reliability of the AMS questions. Data from 335 completed attempts were analyzed to generate Kuder-Richardson reliability and Ferguson's delta statistics which showed the questions to be acceptable. The AMS marking rules were also tested for reliability by calculating Inter-Rater Reliability (IRR) statistics for 8091 question responses. The calculated values of the Cohen's kappa statistic showed that the marking rules were acceptable. This work has demonstrated validity of the AMS with the free-response format, and the reliability of the specific question set posed together with the reliability of the associated marking rules. To demonstrate the applicability of the approach in another domain, the General Relativity Concept Inventory (GRCI), a new free-response concept inventory covering basic General Relativity concepts has been developed and tested. It is concluded that physics concept inventories can be established using free-response questions.

Acknowledgements

I would like to thank Prof Sally Jordan, Dr Holly Hedgeland and Prof Nicholas Braithwaite for supervising my research efforts. In addition, I acknowledge the previous work of Dr Ross Galloway, Dr David Sands and Dr Christine Leach which acted as a precursor to the work which I completed during my PhD. I would also like to thank Moodle developers Tim Hunt and Chris Nelson for helping with my use of the Pattern Match question type. On a related note, I would like to thank the various OU science module teams for supporting my research efforts; in particular, I would like to extend a special thanks to *S217* curriculum manager Michael Watkins for always being on hand to answer queries that I had about using the VLE. Furthermore, I am grateful to the OSL and Isaac Physics platforms for providing suitable places to host the instruments developed during my PhD study. I am also thankful to Prof Andrew Norton, Dr Ulrich Kolb and Prof Robert Lambourne for discussing various aspects of physics and astronomy education with me.

On a personal level, I would like to thank Dr Anita Dawes and Dr Matthew Sylvest for offering continuous advice and support throughout my PhD. On a similar note, I would like to thank Hillary Dawkins who was my PER research partner when I arrived at the OU in 2017. Hillary introduced me to Python programming and taught me about being a grad student. Finally, I reserve a place here to give a huge thanks to two very special people: my younger siblings Lorna and Thomas. Without them none of this would have been possible, and they ever serve as my inspiration and hope.

Contents

1 Introduction

The aim of the research outlined in this thesis is to establish and investigate the use of educational instruments that can be used to test for students' understanding of certain, basic concepts in physics. These instruments make use of technology-enhanced assessment tools, in particular the automated marking of **free-response questions** [1]. The main part of the research focuses on developing a tool which tests for **conceptual understanding** of Newtonian mechanics, known as the *Alternative Mechanics Survey* **(AMS)**; the AMS draws heavily upon years of previous experience from **Physics Education Research**, and adapts a well-known standardized assessment tool to meet its objectives. A secondary part of the research takes these ideas and applies it to the development of a tool to test for conceptual understanding of General Relativity, the *General Relativity Concept Inventory* **(GRCI)**, which is developed *ab initio*.

1.1 Rationale for the research

As currently used in Physics Education Research and similar subject areas, a **concept inventory** is primarily designed as an instrument to learn about student understanding of a specific subject area (Smith and Tanner, 2010). In their current form, most concept inventories make use of the **multiple-choice question** type. Multiple-choice questions are a type of **selected-response question**, meaning that the possible answers are given to the student as a list of options (Jordan, 2013). This format restricts the number of ways in which students can answer the questions and display their understandings and misunderstandings. In contrast, more information about student conceptual understanding would be available if students were to write their own answers in a **constructed-response question** format (Jordan, 2013), and this could be achieved through the use of free-response questions (Rebello and Zollman, 2004). In turn, this would then provide useful information to physics educators, particularly with respect to misconceptions, and so allow them to focus their instruction on these areas.

Making use of free-response questions enables students to give a much broader range of answers than they could with multiple-choice questions, but these free-response answers take much longer to mark by hand than their multiple-choice counterparts.

[1] The first time a term listed in the **Glossary (Appendix I)** is used, it will be **emboldened** in the text.

Therefore, the pretext for this study is that it would be advantageous if the questions on a free-response format concept inventory could be automatically marked. This is in principle possible through technology such as that used in the **PMatch** question type in OpenMark (Butcher and Jordan, 2010), which formed the basis of the **Pattern Match** question type in the **Moodle** question engine. This automated marking also needs to be reliable and accurate, and this can be tested by investigating how closely the automated marking matches with that of expert human markers. Developing automated marking schemes that are on a par with, or better than, corresponding human markers would provide the technology required to take the step from using paper-based multiple-choice concept inventories to using online free-response concept inventories.

Since this is a new and previously untested idea, it is useful to start by adapting an already well-established physics concept inventory into a free-response format version. A suitable concept inventory for this task is the *Force Concept Inventory* **(FCI)** (Hestenes et al., 1992), which is the first and the most well-known of the physics concept inventories (Smith and Tanner, 2010). The traditional multiple-choice version of the FCI has already been extensively used and tested, as summarized by Scott and Schumayer (2017), and is widely agreed to be both a valid and reliable educational instrument. A free-response format version of the FCI is not the same as the traditional multiple-choice version of the FCI, meaning that it needs to be tested for **validity** and **reliability** itself, as it cannot inherit these features from the multiple-choice version of the FCI *a priori*.

The process used to develop the free-response version of the FCI can also be applied to building concept inventories for other subjects. More advanced physics contains more complicated mathematics than Newtonian mechanics, with General Relativity being an example of such a subject. Conlon et al. (2017) conducted an investigation into undergraduate students' ideas about the fate of the universe, illustrating a previous interest in the conceptual understanding of General Relativity. However, there appears to be no General Relativity concept inventory (Stannard et al., 2017). Using the conceptual approach to teaching General Relativity employed by The Open University module *S354 Understanding Space and Time* as a starting point, it has been possible to develop short answer free-response questions to assess conceptual understanding of General Relativity. These questions, when put together, constitute a draft free-response *General Relativity Concept Inventory*. Developing such a concept inventory

is a step towards moving concept inventories into a post multiple-choice question era.

Free-response format concept inventories are fundamentally different from their multiple-choice counterparts, which also means that they could be used for different purposes. Looking ahead, the concept inventories developed in this work could be deployed as teaching tools as well as summative assessment tools. This could be done by giving students **feedback** on their performance, allowing them to see which concepts they have struggled with. The questions could also be used as the basis of discussion in lectures or in small groups, again with the aim of allowing students to investigate shortcomings in their own understanding. Developing concept inventories that can be used in this way would flip the traditional purpose of a concept inventory around; instead of the teaching methods being assessed and appropriately adapted, it could be the students' conceptual understanding itself being assessed, allowing teaching methods to be appropriately adapted in a more direct way.

1.2 Research questions and the approaches used to address them

The research questions which form the basis of the work presented in this thesis are presented below. These are revisited in **Section 10.2** at the end of the thesis, which explains how the work carried out in the thesis has answered them.

• RQ1: To what extent are free-response versions of a physics concept inventory questions valid and reliable?

• RQ2: How reliable are automated marking schemes when used to mark free-response concept inventory questions?

• RQ3: How effective are concept inventories when used to assess the conceptual understanding of a mathematically involved physics subject?

To test for *validity*, a qualitative approach is required. This is because student feedback about the instrument collected in interviews can be used to verify whether it is doing what is was designed to. In this work validity was tested for by conducting **usability testing** with corresponding interviews, a method exemplified by Barnum (2010). Since qualitative data is difficult to interpret in its raw form, the results were analyzed using the technique of **Thematic Analysis** (Braun and Clarke, 2006). Thematic Analysis contains six iterative steps, and is based on the idea of reducing the initial qualitative data set down into the eponymous themes which can be interpreted.

In the context of this study, information learned from the themes allowed investigation of whether the instruments developed were valid.

To test for *reliability*, a quantitative approach is required. This is because the responses given to the questions can be used to see whether the instrument is producing similar results each time it is used. Since the instrument has a free-response format, there were two components to this reliability testing; one strand focused on testing the questions for reliability, whereas the other focused on testing the marking rules for reliability. The entire instrument can only be deemed as reliable when both of these parts have been found to be reliable.

To test the reliability of the questions, the *Classical Test Theory* **(CTT)** approach (Crocker and Algina, 1986) was used. The CTT approach consists of calculating values for five statistics, and checking whether they are within an acceptable range. The CTT statistics calculated in this work were:

• **Difficulty**, which determined how hard each particular item was.

• **Discrimination**, which determined how well each item could differentiate between high-scoring and low-scoring test-takers.

• **Point biserial coefficient**, which determined how well each item aligned with the material tested by the entire test.

• **Cronbach's alpha**, which tested the reliability of the entire test. The commonly used **Kuder-Richardson reliability** formula (Kuder and Richardson, 1937) was used to compute Cronbach's alpha in this work.

• **Ferguson's delta**, which tested the discrimination capabilities of the entire test.

Calculating these CTT statistics checked whether the individual questions on the instrument, and the instrument as a whole, were functioning at an acceptable level. Problematic questions were identified through CTT statistics, and measures were taken to improve these questions, or remove them, as appropriate. Note that a modern test theory, such as *Item Response Theory* **(IRT)**, was not employed in this study because there were not high enough numbers of responses for IRT to give meaningful results (Wallace and Bailey, 2010; Baily et al., 2017).

To test the reliability of the marking rules, the *Inter-Rater Reliability* **(IRR)** approach (Artstein and Poesio, 2008) was used. The IRR approach consists of calculating values for various statistics across a group of markers using a reference set of responses, and checking whether they are within an acceptable range. For this study, the **marking agreement** (Scott, 1955) and **Cohen's kappa** (Cohen, 1960) were the IRR statistics calculated. When two markers are given a set of responses to mark, the marking agreement is the percentage of the cases on which the two markers agree. Markers can however agree out of random chance, and Cohen's kappa calculates the agreement while also taking account of random chance. The marking agreement hence gives an estimate of how often two markers agree with one another; Cohen's kappa then gives a more accurate estimate of what the level of agreement between the two markers actually is.

The IRR statistics were used to test how well the automated marking schemes were marking the responses given by students. This was done by comparing the computer's marking to that of expert human markers; a detailed description of the procedure used is given in **Subsection 6.2.1**. The values of the marking agreement and Cohen's kappa between the computer and expert human marking then showed whether the automated marking schemes were functioning at the required level; where they were not, cases where the computer and human marking disagreed were used to manually improve the computer marking rules.

Concept inventory development is iterative (Porter et al., 2014), and the process of developing effective marking rules for free-response questions is also an iterative process based on responses from users (Jordan, 2009). As a result, several iterations were required to get the free-response format FCI to the required level of functionality for use on a large scale. At each step, responses were gathered, and the CTT and IRR calculations were used to highlight issues with the questions and marking rules, which were modified accordingly. A more detailed outline of this iterative process is given in **Chapter 4**.

The work reported in this thesis was carried out at The UK Open University (OU). The OU was founded in 1969 with the objective of giving an opportunity to pursue higher education to those who had, for whatever reason, missed out on the chance to do so. It is a non-traditional university that delivers its teaching at a distance, and has no entry requirements. OU students, many of whom are studying part-time alongside

employment or other responsibilities, study for various reasons, ranging from earning a degree for their career progression, to studying modules for enjoyment in their free time. Nevertheless, the findings are relevant beyond this distance learning regime. Indeed, much of the testing involved populations of non-OU students.

The OU makes use of and maintains the open-source Moodle question engine, and the concept inventories used in this work were also authored in Moodle. In particular, the free-response questions were authored using the Pattern Match question type in the Moodle question engine; more details about the operation of Pattern Match and how it was used in this work is given in **Chapter 4**. Also within the OU context, the *OpenScience Laboratory* **(OSL)** is the OU's online platform for hosting remote experiments and other activities (The Open Science Laboratory, 2020), and the concept inventories developed in this study were hosted on the OSL for ease of access to participants. In addition, it is relevant to mention that all participants involved in all of the studies outlined in this thesis were volunteers, and the concept inventories were not offered as part of an assessment component of any OU module.

1.3 Outline of the thesis

The rest of the thesis contains the following chapters and appendices:

Chapter 2 reviews the existing literature surrounding the subject of concept inventories and computer marking of free-response questions.

Chapter 3 presents analysis carried out on data collected as a part of this doctoral research using the conventional multiple-choice version of the FCI. This case study illustrates how a concept inventory works in practice.

Chapter 4 outlines the development process used to develop the *Alternative Mechanics Survey* (AMS), which is a version of the FCI that makes use of free-response questions.

Chapter 5 explains the validity testing carried out on the AMS in The Open University's usability laboratory. This data is qualitative, so it is analyzed using *Thematic Analysis*.

Chapter 6 focuses on the reliability testing carried out on the AMS questions and marking rules. *Classical Test Theory* is used to test the AMS questions, whereas *Inter-Rater Reliability* is used to test the AMS marking rules.

Chapter 7 expands the free-response scope of the AMS by putting more of the questions into free-response format. It hence focuses on further testing of the AMS questions and marking rules using the quantitative approaches of *Classical Test Theory* and *Inter-Rater Reliability*.

Chapter 8 extends the use of the AMS by collecting data using a new platform. It is concerned with further development and testing of the AMS marking rules using *Inter-Rater Reliability*.

Chapter 9 applies the ideas used to develop the AMS to develop a concept inventory for another subject within physics. It goes through the steps that were taken to develop a draft version of the *General Relativity Concept Inventory* (GRCI), and this includes analysis of both qualitative interview data and quantitative response data.

Chapter 10 summarizes the work carried out in the thesis, and refers back to the original research questions to highlight how they have been answered. It additionally outlines some possible future directions for the work.

Chapter 11 gives a list of all the references used in the work.

Appendix A gives tables which show how the questions on the different versions of the AMS map to one another, and to the questions on the original version of the FCI. In addition, it also gives the questions and marking rules from the final version of the AMS.

Appendix B gives the questions from the FCI, as used in Chapter 3.

Appendix C gives the questions from Version 1 of the AMS, as used in Chapter 5 and Chapter 6.

Appendix D gives the questions from Version 2 of the AMS, as used in Chapter 7.

Appendix E gives the questions from Version 3 of the AMS, as used in Chapter 8.

Appendix F gives the questions and marking rules from the draft version of the GRCI, as used in Chapter 9.

Appendix G gives the interview questions used in the AMS usability testing study.

Appendix H gives the interview questions used in the GRCI qualitative testing study.

Appendix I is the glossary, and gives a list of specialized terminology and abbreviations used in the thesis.

2 Literature review

2.1 Conceptual understanding of physics

Conceptual understanding of physics has been of interest for at least four decades. The work of Johnstone and Mughol (1976) is an illustration of this. In their work, they surveyed post high school physics students, as well as first-year university students. Their surveys revealed that the following concepts were difficult at both of the surveyed levels:

Motion

- Uniform motion
- Conservation of momentum
- Elastic and inelastic conditions

Energy

- Energy and power
- Heat and temperature
- Latent heat
- Heat transfer

Electricity

- Current, both direct and alternating
- Resistance
- Induction
- Fields
- Electromotive force and potential difference

Other

- Wave motion
- Magnification
- Pressure
- The difference between mass and weight

Among these topics, they identified three main groups for which students had particularly poor conceptual understanding. These were *Motion*, *Energy* and *Electricity*. The findings highlight the conceptual difficulties that students have with physics, even

when it comes to topics that they have covered several times over the course of their studies.

The work of Johnstone and Mughol in 1976 and the work of many other authors serve as a starting point for subsequent investigations into the conceptual understanding of physics. It is relevant to note that topics in Newtonian mechanics (including *Uniform motion*, *Conservation of momentum* and *Elastic and inelastic collisions* in the list above) have long been recognized as conceptual difficulties.

Clement (1982) was interested in students' conceptual understanding of Newtonian mechanics and also in the misconceptions that the students had about the subject. In a qualitative study, Clement video recorded interviews with students in which they solved a variety of Newtonian mechanics problems in a think-aloud setting. Clement found that students frequently had stable misconceptions associated with the common misunderstanding of the relationship between force and acceleration, and that this misconception arises from day-to-day experience. Clement's work illustrates the early use of a qualitative approach to investigation of student understanding in physics, and such approaches are still used today.

Another study from around this time was conducted by White (1983). White gave force and motion problems to 40 high school science students. By analyzing the responses, White reasoned that most of the students had only a partial understanding of Newtonian ideas of force and motion. In particular, students often failed to take into account the initial state of motion of an object when answering questions about its subsequent motion, and frequently failed to determine the effect that an external force would have on an object's speed. This idea is supported by the findings of McCloskey (1983), who also found that students of all levels make use of intuition and Aristotelian thinking when solving problems about force and motion. Both White and McCloskey highlight the prevalence of students making use of incorrect ideas when answering Newtonian mechanics problems.

2.2 An overview of concept inventories

There are many concept inventories used in physics and astronomy. A concept inventory is a multiple-choice research-level instrument that is designed to test students' conceptual understanding of a particular topic (Lindell et al., 2007), with minimal mathematical content. The inventory is based upon a number of key concepts from

the subject (Jorion et al., 2015). For each item in the concept inventory, there are a number of responses, containing one correct answer and other incorrect **distractor** responses. The distractor responses are designed to correspond to common student misconceptions (Sadler et al., 2009). Dick-Perez et al. (2016) observed that the key to constructing an effective concept inventory lies in the selection of appropriate distractors. In their work, designing these distractors was also found to be one of the most difficult aspects of constructing their concept inventory.

The objective of a concept inventory may vary (Smith and Tanner, 2010). Nevertheless, most inventories are designed with the idea of testing the effectiveness of teaching strategies (Porter et al., 2014). For this, students are given the concept inventory to do before instruction takes place, which is known in the literature as the **pre-test**, and are given it to do again after instruction, which is known in the literature as the **post-test**. The pre-test and post-test scores from across the student body are then compared to gauge the effectiveness of the teaching methods used by the instructor (Bailey et al., 2012). The results of this comparison can be used to guide future instruction by addressing misconceptions that appear to have remained. This pedegogical guidance aspect of concept inventories make them popular instruments for use in STEM subjects. Additionally, being able to quantitatively test the effectiveness of teaching interventions makes their use attractive to instructors.

Similarly, there is not one particular method for developing a concept inventory (Reed-Rhoads and Imbrie, 2008), but most concept inventories follow the same rough iterative pattern of development known as the **Delphi process** (Porter et al., 2014). The Delphi process has the following steps:

• Step 1: Gather together concepts that are to be tested using the inventory.

• Step 2: Come up with the questions and responses. This can be done by consulting the literature; using written student essay responses and spoken student interview responses; and by using the judgement of experts.

• Step 3: Administer the pilot test and gather data.

• Step 4: Analyze the data, then return to the previous steps. The fourth step requires careful analysis of the data gathered in third step, and many iterations may be necessary before the concept inventory reaches the desired level of quality.

These steps are a rough guide, however, and concept inventories do not need to follow them strictly in order to be developed and classified as concept inventories. As detailed by Epstein (2013), many concept inventories follow the design and objectives of the first concept inventory, and this particular inventory is the focus of the next section.

2.3 The Force Concept Inventory

Prather (1985) noted the need for a diagnostic tool that could identify students' conceptual understanding of physics topics, and this was taken up by Halloun and Hestenes (1985) when they developed the *Mechanics Diagnostics Test* **(MDT)**. The MDT was designed to test students' ideas about common physical phenomena, and it tested the kinematics concepts of position, distance, motion, time, velocity and acceleration; and the dynamics concepts of inertia, force, resistance, vacuum and gravity. Hestenes, Wells and Swackhamer (1992) further developed the MDT into the *Force Concept Inventory* (FCI), with the idea of the FCI giving a more systematic and complete outline of students' various Newtonian mechanics misconceptions. The FCI tests six Newtonian mechanics concepts; kinematics, Newton's First Law, Newton's Second Law, Newton's Third Law, the principle of superposition and types of forces. The final version of the FCI (Halloun et al., 1995) comprises 30 multiple-choice questions, each designed to have minimal mathematical content.

Although issues with students' conceptual understanding of Newtonian mechanics were previously well documented, there was no way of investigating these without making use of time-consuming qualitative approaches, such as interviews. The FCI was innovative in that it provided a way of learning about student misconceptions through the completion of a short test. However, the FCI was met with a mixture of praise and criticism when it was released. One criticism levelled at the FCI was that it did not measure a single construct (Huffman and Heller, 1995). Using **Factor Analysis**, Huffman and Heller argued that the FCI was a test of mastery of various force-related situations, but was not a test of the force concept itself. Hestenes and Halloun argued back that *Factor Analysis* may not be appropriate for application to the FCI because it requires the total score to be broken down into less meaningful sub-scores. They hence argued that the total score on FCI does indeed measure conceptual understanding of Newtonian mechanics (Hestenes and Halloun, 1995). The debate between the two parties about what the FCI actually measures remains unresolved.

Concerns about the effectiveness of the FCI distractors has also been raised as a possible issue. To this end, Rebello and Zollman (2004) investigated the idea of using free-response questions in a concept inventory. They did this by giving students free-response versions of four FCI questions. They then compared the free-response answers to the corresponding multiple-choice distractors, and looked to see whether the distractors matched up with the incorrect answers given by students in the free-response versions. Rebello and Zollman found that students often gave answers that corresponded to the multiple-choice options given on the FCI, but that there were also cases where students gave answers that did not correspond to any of the options, indicating different thought processes being used to answer the questions. Rebello and Zollman then used these responses to create new distractors for the four FCI questions, and it was found that the new distractors were more effective than the previous distractors when used with other students.

There is also a well-documented gender gap in FCI attainment (Dockter and Heller, 2008), and this gap can be found in the entire test score, as well as on an item by item level (Henderson et al., 2018). This gender gap has been observed at various UK institutions during course instruction, although a similar gap was not present in end-of-term exams where extended responses were required (Bates et al., 2013), leading them to postulate that the multiple-choice question type is one of the factors causing the gap. Traxler et al. (2018) suggested that the gender gap could be caused by eight items on the FCI, of which six have males outperforming females, and two have the opposite trend. By removing these eight items from their analysis, Traxler et al. found that the gender gap effect was halved. The findings of the study conducted by Henderson et al. (2019) suggested that the gender gap is mainly caused by previous physics preparation; this contrasts with the previous work of Madsen et al. (2013), who instead suggested that the gender gap is a result of many different factors, which means that it cannot be easily removed by simply changing the approach used to teach Newtonian mechanics, or by changing the question type on the FCI. Madsen et al. further pointed out that despite extensive analysis, the factors causing the gender gap in FCI attainment are still not properly understood.

Support for the FCI came from Hake (1998) who argued that the use of the FCI alongside **Interactive Engagement** teaching methods could be used to facilitate effective instruction of Newtonian mechanics material. *Interactive Engagement* methods differ from traditional teaching methods in that they make use of a large amount of

student involvement, practical aspects and feedback. An example of such a method is the **Peer Instruction** method of Mazur (1997), where students are encouraged to think independently about a key question and come up with their own answer, before discussing it with their fellow students, and then the instructor. Such approaches led to large-scale reform in physics education (Reed-Rhoads and Imbrie, 2008), with more of a focus on the development and use of new Interactive Engagement teaching methods. These new methods could, in theory, be evaluated for effectiveness by use of the FCI. As a result, the FCI itself is partially credited with much of the reform that has taken place in physics education over the past two decades (Scott et al., 2012).

By 2011, Lasry et al. (2011) estimated that the FCI had been taken more than 100,000 times by students, and it is still used to test the effectiveness of teaching methods (Ding and Caballero, 2014). It is generally agreed within the community that the FCI is capable of measuring something, but there are still disagreements about what this something may be (Wallace and Bailey, 2010). Despite such concerns, the FCI is widely used at a range of institutions, and it is undeniable that the FCI is useful for gathering large amounts of Physics Education Research data.

Scott and Schumayer (2017) analyzed incorrect responses to FCI questions, with the aim of learning about the pre-conceptions of the students answering. They assumed that correlations exist between the items on the FCI, and made use of **Exploratory Factor Analysis** to investigate these correlations. They found that many of the cohort of students ($N = 2109$) made use of coherent non-Newtonian worldviews when giving incorrect answers to the FCI questions. In Scott and Schumayer (2018) the same authors made use of a network analysis approach to look for relationships between the incorrect FCI responses chosen by the cohort. The analysis found that students were not correctly applying Newton's First Law in many cases, which the authors identified as being a major barrier to students attaining a Newtonian way of thinking. This finding was backed up by the work of Eaton and Willoughby (2018), who used **Confirmatory Factor Analysis** to analyze an FCI response data set gathered from 20,882 students.

The FCI itself also remains of interest to physics education researchers. Yasuda et al. (2018) studied cases of answers that were correct but arrived at by faulty reasoning to four FCI questions (Q5, Q6, Q7 and Q16; these questions can be found in **Appendix B**), by giving students follow-up questions to each of the four questions.

The follow-up questions were designed to identify whether the line of reasoning used to give the answer to the first part was correct. They tested 1110 students with these questions, and found that students gave correct answers but faulty reasoning for 11% of Q5 responses; for 57% of Q6 responses; for 62% of Q7 responses; and for 55% of Q16 responses. Yasuda et al. however acknowledged that using sub-questions to test for faulty understanding was a limit of the study, since students could in turn answer these questions correctly using a faulty line of reasoning also. The examples given of Physics Education Research projects based around the FCI serve to illustrate the widespread use that is made of this particular concept inventory.

2.4 Other concept inventories

The FCI's success and subsequent physics education reforms led to the development of concept inventories in other areas of physics and astronomy. There are several examples of such inventories, and the ones presented in this section were chosen because of their relevance to attaining the overall aims of the current research.

In physics, the *Force and Motion Concept Evaluation* **(FMCE)** (Thornton and Sokoloft, 1998) was developed as another concept inventory to test the conceptual understanding of Newtonian mechanics. The FMCE is built around kinematics and dynamics concepts, again using a multiple-choice format, and it contains 47 questions as opposed to the FCI's 30. It has been used with both high school students and students at colleges and universities. Initial findings from deployment of the FMCE found that traditional teaching methods were not effective in removing kinematics and dynamics misconceptions held by students, and these findings led to the development of computer-based active learning curricula. The FMCE serves as an example of a concept inventory being used in the traditional sense of assessing teaching methods, and being used to justify the development of new teaching methods. In addition, the FMCE is an example of a concept inventory that was developed as an alternative to the FCI.

The *Conceptual Assessment Tool for Statics* **(CATS)** (Steif and Dantzler, 2005) was designed to test for the understanding of statics concepts. The CATS contains 27 multiple-choice questions overall, with five categories of question to test various aspects of statics. It was tested at the first-year university level at five institutions, with one institution doing the CATS as a pre-test and a post-test, and the other four running

it only as a post-test. Based on their findings, the authors suggested changes to the use of concept inventories. First, the concept inventory questions themselves could be used to guide conceptually-based instruction; second, concept inventories could be used part way through a semester to guide remedial exercises. Both of these ideas point to a future in which concept inventories could be designed with the intention that they play a more active role in instruction.

The *Colorado Upper Division Electrodynamics Test* **(CURrENT)** (Baily et al., 2017) tests conceptual understanding of advanced electrodynamics. The authors tested the CURrENT for both validity and reliability. To test for validity, the authors conducted student interviews and compared student scores on the CURrENT with their final exam score. To test for reliability, the authors calculated the Classical Test Theory statistics of difficulty, discrimination, Cronbach's alpha and Ferguson's delta. Note that most of the CURrENT questions were posed in the open-ended format in order to directly probe the understanding of the students taking the test. This required the questions to be marked manually, which was a subjective process. As a result, the reliability marking also needed to be tested by calculating the Cohen's kappa Inter-Rater Reliability statistic. The authors found that the open-ended response aspect of the CURrENT was useful for finding about student thinking, although they wished to develop a multiple-choice version of the CURrENT because it would be easier to administer and mark. This work illustrates that open-ended concept inventory questions could provide useful information to physics educators, provided that there was a quick and consistent way of marking them.

The *Brief Electricity and Magnetism Assessment* **(BEMA)** (Ding et al., 2006) was developed to test conceptual understanding of electricity and magnetism. The assessment has 31 questions, and is designed for use with physics students at the introductory college level. This concept inventory differs from other physics-based concept inventories because some of the questions contain equations within the statement of the question and in the available multiple-choice responses. In this way, the BEMA illustrates the difficulties that arise when trying to author conceptual questions for a highly mathematical topic such as Electromagnetism.

The *Relativity Concept Inventory* **(RCI)** (Aslanides and Savage, 2013) was designed to test understanding of topics from Special Relativity. The inventory contains 24 multiple-choice questions based on 9 concepts, and the authors additionally endeav-

oured to select distractor options that matched with common misconceptions identified by the literature. The RCI was designed to measure the change in conceptual understanding of students as a result of instruction, so was used in the same *pre-test* and *post-test* format as the FCI. It is worth noting that the RCI does not include topics from General Relativity.

Bailey et al. (2012) developed the *Star Properties Concept Inventory* **(SPCI)** because there was no such diagnostic tool to test for conceptual understanding of star properties, though there were other concept inventories covering different topics in astronomy. The SPCI was built using the three concepts of stellar properties, nuclear fusion, and star formation. It contained 23 multiple-choice questions based on the scientific content, as well as two contextual questions, and it was used with university undergraduate students on a first-year introductory astronomy course. The SPCI went through several iterations before it was deemed as being ready for use at the large scale. It is also of note that when calculating the Classical Test Theory statistics for the SPCI, only complete attempts were used, in order to avoid skewing the results with low scores which were down to abandonment of the attempt rather than conceptual misunderstandings; this is a typical approach to take when calculating Classical Test Theory statistics, since only complete attempts yield meaningful results in the analysis (Crocker and Algina, 1986). Finally, the SPCI was tested carefully at each iteration, which highlights the interplay between developing a concept inventory and testing it.

2.5 Evaluating concept inventories

This section discusses some key evaluations of concept inventories. Lindell et al. (2007) were motivated by a lack of consistency in development methods to evaluate the development process for twelve physics and astronomy concept inventories. Three classes of resource had been used to determine the concepts being tested: investigations into student understanding; previous literature about student understanding; and researcher understanding. The authors believed the first of these to be the most important, since the concept inventory will eventually be used on students, but only four of the twelve concept inventories made use of investigations into student understanding in their development.

Lindell et al. also investigated the sources used to create the distractors (expert understanding, student understanding, or both) and the correspondence of the distrac-

tors to misconceptions. Again, since the concept inventory will be used with students, it is crucial that student misunderstanding is used to formulate the distractors. Similarly, since the concept inventories will be used to look for gaps in students' conceptual understanding, it is important that the distractors are capable of identifying student misconceptions. Five out of the twelve concept inventories were found to have distractors based on student misunderstanding, whereas only two out of the twelve of the concept inventories were judged to have distractors that could identify student misconceptions.

The authors additionally investigated differences in item statistics, differences in field testing, and differences in establishing validity and reliability. For item statistics, the authors advise that as a minimum, the Classical Test Theory statistics of difficulty and discrimination should be reported; they found that nine out of twelve of the concept inventories reported the difficulty statistic, and eight out of twelve reported the discrimination statistic. For field testing, the authors advise that concept inventories should be tested against large sample sizes, such that statistical analysis returns meaningful results; they found that two thirds of the inventories were used on more than 1000 students, which was an arbitrary benchmark chosen to examine field testing. For validity and reliability, the authors found that none of the inventories were rigorous enough in their validation process, but nine out of twelve of the inventories did enough to call themselves reliable. Overall, Lindell et al. concluded that there needs to be a formal classification for concept inventories, and an agreed standard for the development of such instruments. This would render the development and deployment of concept inventories in STEM subjects consistent with the scientific disciplines that underpin them, a point which has previously been argued by D'Avanzo (2008).

Building on these ideas, Jorion et al. (2015) emphasized the importance of testing for validity and reliability and of using data gathered from actual students when making claims about what a concept inventory can do. The authors proposed an analytic framework for evaluating the claims that are made about concept inventories, and applied this framework to the *Conceptual Assessment Tool for Statics (CATS)*, the *Statistics Concept Inventory* **(SCI)**, and the *Dynamics Concept Inventory* **(DCI)**. The following claims about concept inventories were identified:

• Claim 1: Concept inventory scores can be used to indicate an individual student's overall understanding of all the concepts identified in the concept inventory.

• Claim 2: Concept inventory scores can be used to indicate an individual student's understanding of a specific concept.

• Claim 3: Concept inventory scores can be used to indicate an individual student's misconceptions and common errors.

Jorion et al. contended that these claims can be tested using statistical analysis of the students' performance data. Five statistical analysis techniques were used, with different techniques chosen to target different features of the data: Classical Test Theory (CTT) and Item Response Theory (IRT) were used to test the properties of the items and the tests; Exploratory Factor Analysis and Confirmatory Factor Analysis were used to investigate the structural features of the inventories; and a form of item response analysis called *Diagnostic Classification Modelling* was used to test for student mastery of the concepts and to check the diagnostics quality of the items. The authors found that the CATS and the DCI could be used to test overall student understanding, but the SCI could only partially do this; that the CATS could test for understanding of specific concepts, but the DCI and SCI could not; and that none of the CATS, DCI or SCI was able to measure students' misconceptions and errors.

In light of these findings, the authors suggested that the CATS could be used to measure student understanding of the concepts being tested; on the other hand, the DCI and SCI could be used as low-stakes tests to gauge the most basic level of understanding, but would perhaps be best used as a post-test only since students are unlikely to be familiar with the concepts being tested before they receive instruction. More generally, the study highlighted the importance of rigorous testing of concept inventories to ensure they are capable of doing what they are designed to do.

Smith and Tanner (2010) discussed the function of concept inventories and their possible use in the future. They pointed out that most concept inventories make use of multiple-choice questions, and there is only a limited amount of information that can be drawn from responses given in this format. In this respect, concept inventories are typically *tier-one diagnostics tests*; they only ask students to select an answer, with no further explanation. This issue is addressed in the design of *tier-two diagnostics tests*. In these tests, the student is asked to give an answer, and is also asked to explain their answer. The *tier-three diagnostics test* goes further. In this case, the student is asked to give an answer, to explain that answer, and to also rate the confidence that they

have that their answer is correct. Tier-two and tier-three diagnostics tests show how testing can be advanced beyond simple tier-one multiple-choice to gain extra insight into student thinking.

The authors pointed out that concept inventories are typically given in a pre-test, post-test format, with the scores compared to check for learning gain. The learning gain calculated is then assumed to be down to teaching method alone, although other factors, such as student self-study, are also likely to be relevant. Even if the pre-test and post-test administration format was suitable for the multiple-choice FCI as it was originally designed, there are various scenarios in which making use of a pre-test and post-test format is not appropriate.

In summary, the present design and structure of concept inventories are flawed and, since concept inventories are used to guide teaching, identify misconceptions, test the effectiveness of teaching approaches, and test the conceptual understanding of students, it is important that these flaws are addressed. More positively, Smith and Tanner pointed towards a future in which assessment is not treated as separate from teaching, rather being used to challenge students to articulate their thinking and hence to learn. The concept inventory itself could be cast in this light and used as a teaching tool as well as an assessment tool, as earlier suggested by Chen et al. (2004), who attempted to combine concept inventories with rapid feedback in order to provide real-time self-evaluation learning tools.

Ishimoto et al. (2017) compared results on a concept inventory between different cultures. This is a different aspect of concept inventory evaluation, but is an important consideration to make, since concept inventories are translated into many different languages for use with students of many different cultures, while retaining the same objectives. The authors pointed out that student views of force and motion are influenced by personal experience, so students of different nationalities may have different ideas about force and motion. To carry out the investigation, the authors compared the item response curves for FMCE answers given by Japanese and American students.

It was found that the item response curves for most of the FMCE items were similar for the Japanese and American students, which indicated that students overall have similar ideas of force and motion. Where there were differences, the authors traced these to (i) differences in the Japanese and American education systems; (ii) translation

of certain questions from English into Japanese altering the question meaning; and (iii) possible differences in personal experience arising from their respective different cultures. The authors concluded that, whilst in most cases educational practice and teaching interventions may be shared between Japanese-speaking and English-speaking cultures, in other cases, different educational practice and teaching interventions may be required on a case-by-case basis. They highlighted the need for further research in this area, involving other languages and cultures.

Laverty and Caballero (2018) evaluated the FCI, the FMCE, the BEMA and the CSEM using a three-dimensional teaching and learning model that they had developed. The authors found that these commonly-used concept inventories were unable to sufficiently test for student learning as outlined by their three-dimensional model. One of the reasons that they identified for this was that these concept inventories make use of multiple-choice questions, which involve students selecting answers from a pre-prepared list of responses rather than having to write their own. The authors hinted at future concept inventories abandoning the multiple-choice model, in line with the suggestions of Rebello and Zollman (2004) and Smith and Tanner (2010), but this would make the administration and marking of such inventories difficult. The following section explores the use of multiple-choice questions and their limitations in more detail.

2.6 Multiple-choice questions

In order to test for conceptual understanding of physics, multiple-choice concept inventories are commonly used. The earliest multiple-choice questions were used to test US army recruits in World War I (Mathews, 2006), whilst muliple-choice questions were first used in the educational context later in the 20th century (Bacon, 2003). An advantage of multiple-choice questions is that they are versatile, meaning they can be used to test a wide range of topics (Bull and McKenna, 2004); on the other hand, free-response questions are advantageous because they can be used to test precise content (Betts et al., 2009; Ferrao, 2010). Because of the way that multiple-choice questions are set up, it is difficult to make use of them to assess more advanced learning outcomes (Conole and Warburton, 2005), although there have been some efforts to do so (Itza-Ortiz et al., 2003; Gwinnett and Cassella, 2011). The main motivation for using multiple-choice questions over other question types has been that they can be marked quickly, which makes a practical difference to academics who have large class sizes

to assess (Woodford and Bancroft, 2005). However, the same authors asserted that the quality of the multiple-choice questions needs be high to justify their use, which requires such questions to be carefully designed and tested.

A common flaw identified with multiple-choice questions is that it is sometimes possible to figure out the correct answer by working through each of the answer options to seek a *least implausible option*, meaning that such questions are not assessing the skills that they are supposed to (Sangwin, 2013). Additionally, it is possible for students to give correct answers to multiple-choice questions by guessing; when this happens, the student is not demonstrating the intended understanding (Crisp, 2007). A counter-argument to this criticism is that it is unlikely that a student could pass an entire multiple-choice test through guessing alone (Downing, 2003); however, it stands to reason that well-placed, successful guesses on multiple-choice questions can make a difference in the grade awarded to a student in a *borderline* case (Burton, 2005). In addition, there are different cognitive processes involved in answering free-response questions and multiple-choice questions (Nicol, 2007). With this in mind, there are ways in which the effectiveness of multiple-choice questions can be increased, such as by developing plausible distractor options (Dick-Perez et al., 2016), or by carefully considering the order in which the questions are asked (Gray et al., 2002). To ensure that questions are able to test the understandings that they are designed to, it is important to prepare high-quality questions, regardless of whether they are multiple-choice or free-response (Bull and McKenna, 2000).

Simon and Snowdon (2014) compared multiple-choice and free-response questions in the context of assessing coding skills. Students were given a selection of multiple-choice and free-response questions based around interpreting code, and it was found that students in general found the free-response questions to be the more difficult of the two questions types. The authors pointed out that with multiple-choice questions, the student needs to choose from a list of ideas constructed by somebody else, and can also employ guessing and elimination strategies to reach the answer; on the other hand, the authors postulated that free-response questions require a student to think more deeply, since they are required to construct their own answers, and that further insight about student thinking can be gained from students free-response answers as a result of this.

A comparison of user response to multiple-choice and free-response questions was undertaken by Bridgeman (1991). Bridgeman used the *Graduate Record Examination General Test (GREGT)*, a standardized test taken by applicants to US graduate schools, as the basis for the study. A set of students were given multiple-choice versions of some of the GREGT questions to answer whilst a control set students were given free-response versions of the same questions to answer. Correct and incorrect answers from the multiple-choice version of the questions were compared to the answers given to the free-response version of the questions, and it was found that the scores on the multiple-choice questions and the corresponding free-response versions were similar. Bridgeman concluded that, under certain circumstances, multiple-choice questions and free-response counterparts can be used interchangeably to assess student understanding.

Another study comparing multiple-choice and free-response questions was carried out by Funk and Dickson (2011), who aimed to investigate the effect of these question types on student performance. To this end, the authors created multiple-choice and free-response versions of a set of psychology exam questions. These questions were split and given to fifty students to attempt as follows: twenty-five students did 10 free-response questions before doing a 50-question multiple-choice test; the other twenty-five did 10 free-response questions after doing a 50-question multiple-choice test. In both cases, performance on the multiple-choice versions of the questions was higher than on the free-response versions, indicating that the students found the free-response variants of the questions to be more difficult than their multiple-choice counterparts. This finding agrees with that of previous work by Hudson (2010), who found that multiple-choice questions were easier to answer than free-response questions for high school chemistry students.

Woodford and Bancroft (2005) pointed out that the apparent difficulty of a multiple-choice question can reveal flaws in its construction: if a question appears easy then its distractors may be ineffective; if it appears difficult then the distractors may be misleading, or the question wording may be ambiguous. Whatever the difficulty reveals about the question, appropriate steps must be taken to improve it if there is an issue, as outlined in Woodford and Bancroft's design priorities. On the test level, the number of questions in the test overall, as well as the order of these questions, will affect the effectiveness of the test. On the question level, the number of options given to the multiple-choice question, the order of these options, the wording of the question itself

and the wording of the distractors all affect the effectiveness of the question. Woodford and Bancroft argued that changing these factors on a test-level and a question-level basis can lead to improved quality multiple-choice questions.

Woodford and Bancroft proposed that provided their design priorities are applied, multiple-choice questions can be used to test for a deeper level of understanding, but the examples that are given refer to mathematical questions. However, it can be argued that multiple-choice questions, no matter how refined or well-designed, may be inappropriate for testing conceptual understanding (Conole and Warburton, 2005), which is more descriptive and less procedural.

Shuhidan et al. (2010) gave questionnaires to 66 computer programming instructors to investigate their perspectives of using multiple-choice questions to assess their students. Over one third of the instructors had confidence in using high quality multiple-choice questions to assess student learning outcomes. Instructors would choose to use multiple-choice questions with their students because they can be answered by weaker students, they are good for revision, and because they prepare students for multiple-choice questions in other courses. On the other hand, some instructors were discouraged from using multiple-choice questions because they felt that they were too easy, and that they encouraged students to guess. Of these instructors, some favoured the use of free-response format questions, both essay and short-answer lengths, to test student understanding.

There also appear to be differences between how males and females perceive and perform on multiple-choice questions, though the reasons for this are not clear (Ben-Shakter and Sinai, 1991). Gipps and Murphy (1994) found that 15-year old girls preferred answering free-response questions, whereas 15-year old boys preferred answering multiple-choice questions. Ben-Shakter and Sinai (1991) found that males outperformed females on multiple-choice questions for both the ninth-grade and university applicant age groups. Ben-Shakter and Sinai did not believe that differences in guessing patterns between males and females were alone responsible for differences observed in performance on multiple-choice questions. This differs from the findings of Richardson and and O'Shea (2013), who found that males and females were as likely as each other to get a question right when they attempted it, but that males were more likely to make an attempt in the first place. Such discrepancies in the literature highlight the poor understanding of demographic differences in responses to assessment

items.

The literature presented in this section has suggested that multiple-choice questions are versatile and easy to use, but they are inappropriate for assessing advanced learning outcomes, such as conceptual understanding. In addition, students can employ eliminate-and-guess strategies to work out the correct answer to multiple-choice questions, or simply guess the correct answer without any knowledge of the subject matter being tested. Further, multiple-choice questions require students to select an option from a list of responses constructed by somebody else, so they are not required to use their own thinking to answer the questions. Use of a free-response question format would encourage students to think deeply, and the answers to such questions would provide useful insight to educators about student thinking. However, free-response questions are not straightforward to mark and are therefore time-consuming for human markers, which multiple-choice questions are not. If there were a way of automatically marking free-response answers, then this could provide a viable alternative to the widely used multiple-choice questions. The idea of automatic marking is the focus of the next section.

2.7 Automated marking of free-response questions

More information about student understanding is available through free-response answers, but these are known to take a long time to mark by human markers, especially for large class sizes. In order to mark free-response answers rapidly, an automated approach where a computer does the marking is required. In addition, the reasoning required to mark such questions effectively is a somewhat subjective process. In order to mark free-response answers effectively, this automated approach must be as accurate as human markers (Jordan and Mitchell, 2009).

Perez-Marin et al. (2009) reviewed the literature of the time on the automatic marking of free-response answers. They reviewed the state of the field at the time, rather than trying to introduce new ideas and techniques. They pointed out similarities in the approaches developed by different groups, as well as discussing the ways in which different authors had attempted to evaluate and validate their methods. They indicated as a summary that the field had greatly developed in the small time that it had been around as a viable line of study, but that there was still much to do before an ideal automated marking system for free-response answers could be found. However,

they took the widespread use of automated marking schemes as an indicator that such systems were on the right track for the development of an ideal marking system.

Some of the earlier attempts to develop the automated marking of free-response questions made use of the *Bilingual Evaluation Understudy* **(BLEU)** algorithm, which was developed at IBM (Papineni et al., 2001). Initially developed as an algorithm to allow machines to translate between different languages, it was adopted by others to perform different tasks. Perez et al. (2004a; 2004b) applied the BLEU algorithm to mark short essay responses written by students. This version of the algorithm was based around the identification of keywords for marking, and did not employ any further constructs beyond this. The authors checked the algorithm's marking against an experienced person's marking, which is known as checking the *marking agreement*, and is a standard measure used to test automated marking. Two of the same authors advanced the work (Alfonseca and Perez, 2004) by using several variants of the BLEU algorithm to mark student essays by combining the BLEU algorithm with shallow *Natural Language Processing* **(NLP)**.

NLP is a field of study that examines the interaction of computers with human language (Pereira and Grosz, 1994). It is required because human language is significantly different from computer language. In particular, the grammar used and the large number of unspoken and inexact ways in which people communicate with one another are different from the way in which a human interacts with a computer, which must follow strict grammar rules of the computer language and must hence be exact. The objective of NLP is to try and find ways to input human language into computers in a way that the computers can understand it.

Alfonseca and Perez also collaborated with others in an attempt to advance the automated marking of free-response answers by combining the BLEU algorithm with *Latent Semantic Analysis* **(LSA)** (Perez et al., 2005). Building upon their previous work, the authors compared their BLEU-inspired marking algorithm with another algorithm based on LSA. Briefly, LSA is a Natural Language Processing technique that operates by finding links between the meaning of a text and the words in that text. The authors also attempted to combine the two approaches, and tested this as well. The BLEU-inspired algorithm was found to be more effective at marking the free-response answers than the LSA algorithm, but the combined algorithm was found to outperform both of the individual approaches. The authors believed that their

study illustrated that it was possible to integrate LSA and other NLP techniques into automated marking algorithms, and they wished to see more advancement in the field of applying automated marking to computer assisted assessment in the future.

Noorbehbahani and Karden (2011) used another modified version of the BLEU algorithm to mark free-response answers. In order to build the automated marking rules, the authors made use of real student free-response answers. These were used to construct a bank of reference answers, and answers were then scored based on similarity to these. As is standard procedure for building effective rules, many student answers were required, and these needed to be manually marked in order to produce the initial reference answer bank to train on. However, if the automated marking worked well, this manual marking step only needed to be done once.

Leacock and Choddorow (2003) developed a scoring engine to mark short free-response answers, which they called the **C-rater**. The authors pointed out that there are many ways in which a concept can be expressed in natural language, so the C-rater marking engine needed to be able to tell when a concept was being expressed and when it was not. To approach this problem, the authors viewed correct responses as being paraphrased versions of a model correct answer; hence they designed the C-rater to be a paraphrase recognizer, instead of being a word recognition tool.

The primary task of the C-rater was to recognize equivalent meaning in short answers. It started with a model correct answer, and the goal was to map student responses onto the model answer, determining whether the student answer was correct or incorrect in the process. The model answer was constructed by hand, but the mapping of the answers was automated. Notably, because the C-rater made use of a single correct answer, it could not be used to deal with open-ended questions; instead, it could score questions that were looking for specific ideas, and that had a definite correct answer.

The C-rater did more than simply match strings of words as a means of marking questions, as it made use of other features as well. These features included analyzing the structure of the answer, looking at the argument followed and also for the presence of pronouns and synonyms. These extra features were considered because the C-rater's approach to automatically marking questions involved the steps of parsing the answer so that the computer could understand it, which allowed a mark to be assigned based on a comparison between the answer and a set of marking rules.

In terms of use and testing, the C-rater was used to score 100,000 responses to 11th grade reading comprehension as well as to score 250 responses to the *National Assessment for Educational Progress* exam, which is a standardized exam taken by school students in the US to check their overall progress. In both cases, the C-rater was found to agree with human marking 84% of the time. The authors made use of marking agreement as well as the Cohen's kappa statistic to assess the effectiveness of their automated marking engine. Significantly, the authors envisioned two possible uses for the C-rater: as an assessment tool, through the use of its automated marking facility; and also as an instructional tool, as instant feedback can be given once the answers have been marked.

Sukkarieh et al. (2003) pointed out that many types of academic exam have questions which require short free-response answers of a few words or sentences. These questions may have corresponding instruction words such as state, suggest, describe and explain. There are many such questions, and they do not require a high amount of cognitive strain to mark, but they do take time to mark. As a result, the authors realized that an automated system that could fully or partially mark such responses would be useful from a marking efficiency standpoint. However, at the time, it was thought that such a system would be too difficult to fully develop.

Pulman and Sukkarieh (2005) continued on from their 2003 work, trying to use computational linguistics methods to mark short-answer free-response answers. The 2003 work made use of an **Information Extraction (IE)** process, which requires marking rules to be authored by hand. The authors pointed out that this is problematic because it requires question authors to have expertise in both the subject domain and in computational linguistics, the latter of which most instructors in non-computing disciplines would not have. Automating the rule creation process would hence be desirable. In the 2005 work, the authors tried employing **machine learning** approaches such as *Inductive Logic Programming* **(ILP)**, **Decision Tree Learning** and **Naive Bayesian Learning** to automatically generate marking rules. Each of the methods showed some promise, but none of them was capable of generating rules that marked accurately enough for them to be used in place of the previously developed IE approach. This work illustrated that the problem of automatically authoring marking rules was still an open problem, and this is still the case today (2020). Additionally, the larger over-arching problem of automatically marking free-response answers remained an open problem, despite these advances.

Mitchell et al. (2002) devised a software based on computational linguistics to try and automatically mark short free-response answers. The resultant **AutoMark** software, developed by *Intelligent Assessment Technologies Ltd.* **(IAT)**, operated by matching free-response answers to sentence templates. Jordan and Mitchell (2009) made use of this software to investigate the process of developing marking rules for free-response questions. Seven free-response questions were authored, and a set of IAT marking rules was written for each question. These questions were designed to assess objective constructs, and to have a small number of feasible correct answers. It was found that using real student responses was important in the rule development and testing process, and it typically took several hundred responses to develop effective marking rules for each question. Jordan and Mitchell reasoned that the time investment required to develop the computer marking was worthwhile, because it would save time for teachers in the future; this time could subsequently be invested into marking more complicated questions, or in supporting students.

Butcher and Jordan (2010) built on this work by comparing the computational linguistics-based approach of the IAT software to two algorithm-based systems; these two systems were the OpenMark *PMatch* question type (Butcher, 2008) and **Regular Expressions**. PMatch is a word matching algorithm that searches for keywords and accounts for synonyms and misspellings, whereas *Regular Expressions* is a string-search algorithm; neither makes use of grammar or syntax in their algorithms. PMatch and Regular Expressions were used to author marking rules for the same seven free-response questions as used in Jordan and Mitchell (2009), and it was found that these marking rules were capable of performing at a similar level to those authored using the IAT computational linguistics approach.

Butcher and Jordan discussed issues with various approaches to marking. For human markers, the quality of the marking is dependent upon the clarity of the marking guidelines. In addition, humans are inconsistent in how they deal with partially correct or incomplete answers, and can also make random mistakes while marking. For the IAT's computational linguistics approach, correct and incorrect answers can be missed because they are not programmed into the mark scheme template, and the spell-checking feature can also fail when specialized vocabulary is used. Similar computer marking issues can also be found in algorithm based approaches such as PMatch and Regular Expressions, where marking rules missing synonyms for the correct answer can lead to **false negative** answers, and marking rules that miss negations can lead to

false positive answers. Despite these issues, Butcher and Jordan repeated the view previously expressed in Jordan and Mitchell (2009) that the development of automated marking is a worthwhile task because of its time-saving benefits. In addition, Butcher and Jordan speculated about automating the process for authoring marking rules, in order to make this easier for potential users.

Building on the work of Butcher and Jordan (2010), Willis (2010) developed an automated marking scheme based around the principles of machine learning. The aim of the research was to see if the large amounts of student response data available could be used to automatically generate the same sorts of mark schemes that had previously been authored manually. A logical approach was chosen instead of a statistical approach because the logical scheme is able to show why a certain response has been marked as right or wrong. The work builds on the previous work of Butcher and Jordan (2010), by developing an algorithm for the PMatch free-response question type.

The author translated the PMatch pattern description into the *Prolog* programming language (Sterling and Shapiro, 1994), and then used the *Inductive Logic Programming* (ILP) system *Aleph* (Srinivasan, 2004) to automatically induce the mark schemes (Muggleton and de Raedt, 1994). The system was given four types of input to induce the marking schemes:

- A set of positive examples - the answers that were marked as correct by an expert human marker.
- A set of negative examples - answers that were marked as incorrect by an expert human marker.
- A hypothesis language - the marking rules.
- Background knowledge - the text of the responses.

These inputs are all annotated because ILP is a form of supervised machine learning. The marking of the ILP scheme was compared against the results from the PMatch and IAT systems tested in Butcher and Jordan (2010). The marking agreement was found to be consistent across all three systems, meaning that short free-response answers can be accurately marked using pattern-matching techniques, such as PMatch and ILP; and they can also be accurately marked by using a deeper syntactic analysis scheme, such as IAT. The language required for pattern matching was also found to be simple enough such that established machine learning techniques could be used to

learn the mark schemes from the correct and incorrect responses. There were also instances where the ILP scheme generated counter-intuitive rules. For some of these, the rules generated gave too much coverage, which led to instances of false positives because the rules marked some of the *incorrect* answers as *correct*. For others, the rules generated were obsolete, since they were already covered by other rules. These cases illustrate the need for a subject specialist to check and refine the rules as necessary.

Issues were raised about what style of questions are suitable for automated assessment. Answers which contained both right and wrong aspects were difficult to mark, and some questions inadvertently encourage students to give these sorts of answers. There were difficulties with marking questions that required specialized language (including chemical equations and mathematical equations), pointing to the possible need for an additional functionality within the marking software to handle such cases. In spite of these issues, Willis concluded that automatic learning can still drastically reduce the amount of time that it takes to author marking schemes for short answer free-response questions; this is of use to educators who wish to make use of these questions as part of their teaching and assessment.

Klein et al. (2011) attempted to develop an automated marking system based on the principles of LSA and by grouping data together through clustering techniques. The authors wanted the system to be used in a summative assessment scheme, so one of its main objectives was to be as accurate as possible. The system is based on an algorithm that is designed to mark free-response answers that are much shorter than essay-length responses, and as a result, it focused on content rather than structure. The system needed to be configured by varying the parameter values controlling the automated marking in order to give more accurate marking results. There was no viable way to automate this process, so it was carried out manually, which was time-consuming. With the right configuration, the system was capable of marking consistently when compared with corresponding marks given by human markers. The authors felt that if there were some way of quickly finding the appropriate parameters for the configuration, then the system could provide a solution to the problem of accurately marking short free-response answers. Based on their own progress, they expected that the automated marking of short free-response answers would one day be possible.

Zehner et al. (2016) tried to develop an automated marking scheme using clustering and machine learning techniques. They pointed out that it was important when using any machine learning scheme to try and avoid training the machine too closely to a particular data set, as this can lead to **over-fitting** whereby the algorithm can operate well on the data set from which it was trained from, but cannot operate effectively on other data sets. They reflected that once the automated marking scheme is developed, it can be integrated into other schemes, so that any costs going into its development can be quickly balanced out by the benefits. The authors recognized a need for a fully-automated approach to developing computer marking rules, as their own approach involved some manual input in its setup.

Cuff et al. (2019) reviewed methods used to automatically score essays and short answer free-response questions, and they pointed out that the approaches used to automatically mark these two questions types are not the same. Their review found that there were many different approaches used to mark short answer free-response questions, and that many of these methods would not be familiar to a non-specialist user. The authors further linked this to the idea that computers are unable to understand the answers in the same way as humans, and that such facets of user perception will be important in determining whether automatically marked questions are to be widely used in the future. Cuff et al. concluded by stating that more research and development is needed into automatic marking in order to make the approach viable, but that it could ultimately be used to improve the reliability of marking in a cost-effective way in the future.

2.8 Summary and looking ahead

Chapter 2 has reviewed the literature pertaining to concept inventories and free-response questions. It found that concept inventories in physics have become an established element for a number of well-recognized conceptual challenges, although technical/mathematical procedural challenges are generally outside the scope of concept inventory methods. However, the utility of surveys of conceptual understanding is compromised to a degree by the marking burden, and this can be ameliorated by recourse to automated marking schemes. The literature pertaining to this area has been explored, highlighting recent advances in the use of **Artificial Intelligence**.

Chapter 3 illustrates how the multiple-choice FCI can be used in practice to investigate student misconceptions, and to assess the effectiveness of physics teaching.

3 Case study of the operation of a multiple-choice concept inventory

3.1 Rationale

The literature review detailed a few of the numerous published concept inventories, which are used to evaluate the effectiveness of teaching methods through a *pre-test*, *post-test* administration methodology. Most of these concept inventories are comprised of multiple-choice questions. The aim of the current research is to develop new concept inventories that make use of free-response questions, so it makes sense to start by examining how a well-established concept inventory operates. An example of such a concept inventory is the FCI itself, and the case study presented here is based on data collected using the standard multiple-choice FCI.

3.2 Methods

3.2.1 Data collection

The Open University (OU) was one of the higher education institutions involved in the Institute of Physics' *Expanding Conceptual Understanding in Physics* **(ECUIP)** project. The ECUIP project was a collaborative project that aimed to assess and improve physics teaching and learning at the participating higher education institutions. In order to gather the required educational data, participating universities administered the FCI as a pre-test and post-test over several academic years.

At The OU, the students taking the level two (OU level two is equivalent to the UK Framework for Higher Equation Qualifications level five) module *S217 Physics: From Classical to Quantum* were selected to take part in the study. *S217* covers core physics topics such as Newtonian mechanics, meaning that the FCI fits naturally with the study materials. As *S217* is delivered completely online, the FCI also needed to be presented online for the ECUIP data collection. As a result, the FCI was put into an electronic form using the Moodle question engine, and offered to students as an optional activity on the OpenScience Laboratory. The pre-test version of the FCI was offered under the name of *Mechanics Survey*, and the post-test version was offered under the name of *Repeat Mechanics Survey*, in order to prevent students from searching for the correct answers to the FCI questions online. Students completed these surveys as optional activities on their own computers and in their own time. In addition, scores

on the surveys did not contribute to students' *S217* module scores, and there was no additional incentive offered for participation. Data were collected in the academic year 2016-2017, and it consisted of *S217* students' scores on individual questions, as well as their total score on the entire test. Not every student who did the pre-test did the post-test, and vice-versa.

3.2.2 Data analysis

Difficulty

Difficulty is defined to be the proportion of test-takers who answer the question correctly (Crocker and Algina, 1986). Difficulty is hence a measure of how easy a question is. For instance, a difficulty of 0.9 would mean that 90% of the students got the question right, making this an easier question than one with a difficulty of 0.1, where only 10% of the students got the question right. Questions which are too easy or too difficult are unable to discriminate between students at different performance levels, so a difficulty value of around 0.5 is preferred for individual questions (Ding and Beichner, 2009). In practice, it is not feasible to design every item on a test to have a difficulty value of 0.5. As a result, difficulty values within the range [0.3, 0.9] are acceptable, because this eliminates the items which are abnormally easy or hard (Doran, 1980).

Normalized gain

Normalized gain is designed to be a measure of learning gain between a pre-test and a post-test, and the higher the normalized gain, the greater the learning gain is. Two types of normalized gain can be calculated; the mean of the normalized gains of each student, and the normalized gain of the mean total scores. Both give an indication of whether the students have increased or decreased in their understanding as a result of studying the material. It hence gives an overview of the progress of the student cohort as a whole.

For an individual student, the normalized gain g is calculated using the following formula (Physport, 2018):

$$g = \frac{post - pre}{100 - pre} \tag{3.1}$$

Where *pre* is the student's pre-test percentage score, and *post* is the student's post-test percentage score.

Hence the mean of the normalized gains $\bar{g}$ is calculated using the following formula:

$$\bar{g} = \frac{\Sigma g}{N} \tag{3.2}$$

Where Σg is the sum of all of the students' individual normalized gains, and N is the total number of students who did both the pre-test and the post test.

In addition, the normalized gain of the mean scores $\langle g \rangle$ is calculated using the following formula (Hake, 1998):

$$\langle g \rangle = \frac{\langle post \rangle - \langle pre \rangle}{100 - \langle pre \rangle} \tag{3.3}$$

Where $\langle pre \rangle$ is the mean pre-test percentage score, and $\langle post \rangle$ is the mean post-test percentage score. Note that only the scores of students who completed both the pre-test and post-test are used in this calculation.

For FCI scores, Hake (1998) defined the following ranges for normalized gain calculations:

- *Low-g* where $g < 0.3$.
- *Medium-g* where $0.3 \leq g < 0.7$.
- *High-g* where $0.7 \leq g$.

Normalized change

The **normalized change** c is a modified normalized gain which takes into account the effect of students performing particularly highly on the pre-test. There are two types of normalized change calculated; the mean of the normalized changes of each student, and the normalized change of the mean total scores. As for the normalized gain, both give an indication of whether the students have increased or decreased in their understanding as a result of studying the material. It hence gives an overview of the progress of the student cohort as a whole.

For an individual student, normalized change c is calculated using the following formula (Marx and Cummings, 2007):

$$c = \begin{cases} \frac{post-pre}{100-pre} & \text{if } post > pre \\ \text{drop} & \text{if } post = pre = 0 \text{ or } 100 \\ 0 & \text{if } post = pre \neq 0 \text{ or } 100 \\ \frac{post-pre}{pre} & \text{if } post < pre \end{cases} \tag{3.4}$$

Where *pre* is the student's pre-test percentage score, *post* is the student's post-test percentage score, and *drop* indicates that the calculation is not used. The rationale behind the calculation is that those scoring higher on the post-test will have a positive c; those scoring lower on the post-test will have a negative c; those who score the same on the pre-test and the post-test (but not 0% or 100%) will have a c value of zero; those who score 0% on both the pre-test and the post-test are not used in the calculation to prevent the results from being skewed by uncharacteristically low scores; and those who score 100% on both the pre-test and the post-test are not used in the calculation to prevent the results from being skewed by uncharacteristically high scores.

Hence the mean of the normalized change $\bar{c}$ is calculated using the following formula:

$$\bar{c} = \frac{\Sigma c}{N} \tag{3.5}$$

Where Σc is the sum of all the students individual normalized changes, and N is the total number of students who did both the pre-test and the post-test.

Further, the mean of the normalized change $\langle c \rangle$ is calculated using the following formula:

$$\langle c \rangle = \begin{cases} \frac{\langle post \rangle - \langle pre \rangle}{100-\langle pre \rangle} & \text{if } \langle post \rangle > \langle pre \rangle \\ \text{drop} & \text{if } \langle post \rangle = \langle pre \rangle = 0 \text{ or } 100 \\ 0 & \text{if } \langle post \rangle = \langle pre \rangle \neq 0 \text{ or } 100 \\ \frac{\langle post \rangle - \langle pre \rangle}{\langle pre \rangle} & \text{if } \langle post \rangle < \langle pre \rangle \end{cases} \tag{3.6}$$

Where $\langle pre \rangle$ is the mean of the pre-test percentage scores, $\langle post \rangle$ is the mean of the post-test percentage scores, and *drop* indicates that the calculation is not used. The rationale behind Equation (3.6) is the same as that for Equation (3.4) above.

The properties of the normalized change calculation provides it with some advantages over the normalized gain. One advantage is that it removes the low pre-score bias present within the normalized gain calculation, whereby it is easier to have a higher

normalized gain value when the pre-test score is lower. Secondly, a perfect pre-test score yields an unbounded normalized gain value, which is avoided by the normalized change because c takes a value between -1 and 1. Taken together, this means that the normalized change can be meaningfully calculated in a wider range of situations than the normalized gain, making its use preferable.

Marx and Cummings (2007) pointed out that each individual diagnostic instrument would have its own appropriate ranges for *Low-c*, *Medium-c*, and *High-c*. Both g and c are able to take negative values in the event of a lower post-test score, and have an upper value of 1 for higher post-test scores. As a result, taking Hake's previous definitions for FCI normalized gains as a guide (Hake, 1998), the following ranges for normalized change calculations are used in the current study:

- *Low-c* where $c < 0.3$.
- *Medium-c* where $0.3 \leq c < 0.7$.
- *High-c* where $0.7 \leq c$.

Before presenting the findings from the study, it is worth recalling that the FCI pre-test and post-test were offered as optional activities to students in the 2016-2017 presentation of the *S217* module, meaning that not every student on the course engaged with the pre-test or the post-test. It is likely that only the most enthusiastic students engaged with these activities (Hunt and Jordan, 2016), and these students are often also the most able ones, which will have affected the calculated statistics through a selection effect. The reliability of the data could have been improved by making the activities compulsory and also had it been possible to offer them to students during a face-to-face class; however the distance-learning nature of Open University teaching made this impossible. This is acknowledged as a limitation of the study, and it is referred to where relevant in **Section 3.3**.

Note that in what follows, the FCI question numbering used is that of the standard multiple-choice FCI (Halloun et al., 1995). The FCI questions as given on the OSL to conduct this study can be found in **Appendix B**. In addition, the mapping of these questions to those on the different versions of the free-response AMS developed in this doctoral research project are given in Tables A.1 and A.2 of **Appendix A**, and the corresponding AMS versions of the questions can also be found in **Appendix A**.

3.3 Results and Discussion

In the academic year 2016-2017, 65 students completed the FCI pre-test and 28 completed the FCI post-test. Of these, 27 students completed both the pre-test and the post-test. These results are considered here, initially as distinct data sets, and key findings are discussed as they are raised.

3.3.1 Findings from the 2016-2017 FCI pre-test

The frequency of each of the different scores for the $N = 65$ completed pre-tests are shown by the blue bars in Figure 3.1 below. Note that the frequency of each of the different post-test scores (given by the orange bars) are also shown for comparison.

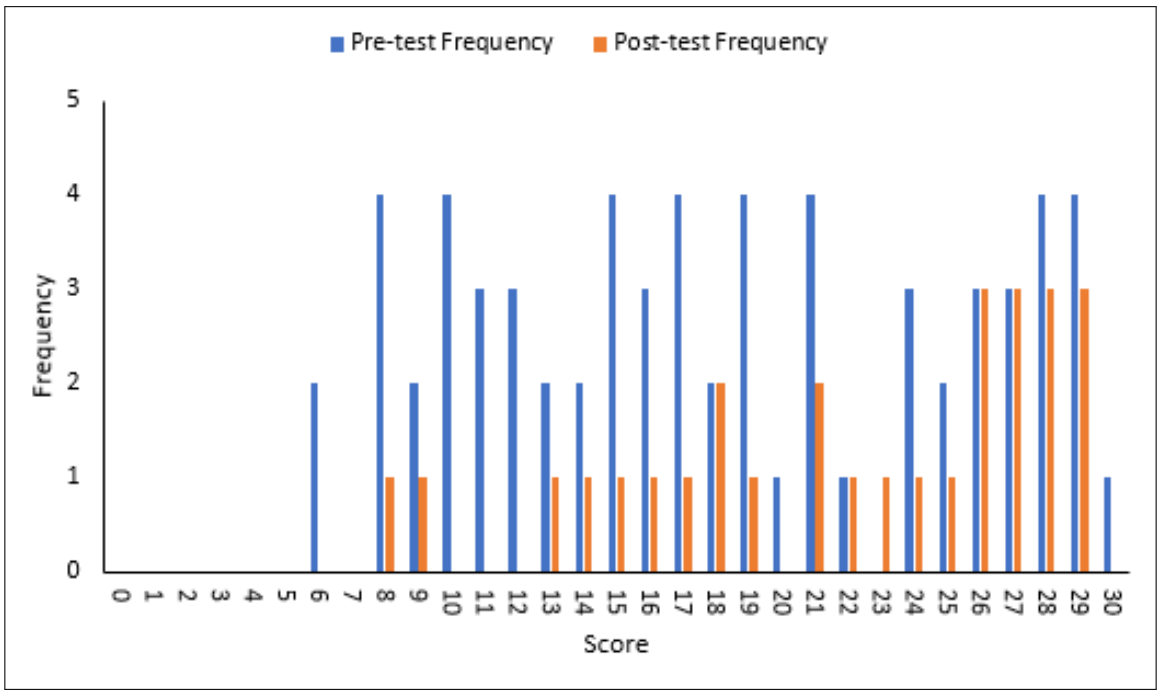


Figure 3.1: Graph showing the distribution of the 2016-2017 FCI pre-test and post-test scores for all 60 completed pre-tests, and all 28 completed post-tests.

For the pre-test data, there is no clear pattern from the shape of the graph. The mean score was 18.12 and the median score was 17. The lowest score attained was 6, whereas the highest core was 30, which is also the maximum possible score on the FCI. Note that the FCI contains 30 questions with 5 multiple-choice responses each, so the expected score for a student randomly selecting responses by guessing (known in the literature as the *random guess score* (Draaijer et al., 2018)) would be 6.

It is also useful to look at student performance on the individual items on the test. To do this, the difficulty of each of the questions was calculated for the 2016-2017 pre-test, and the results are shown in Figure 3.2 below, along with the concept that each question tests understanding of. The typology used to classify the FCI questions by concept was designed by the author and a member of the supervisory team. The system uses five concepts out of the six that the FCI was originally designed to test (Hestenes et al, 1992) because the sixth concept, *Superposition of Forces*, is tested implicitly whenever forces are combined.

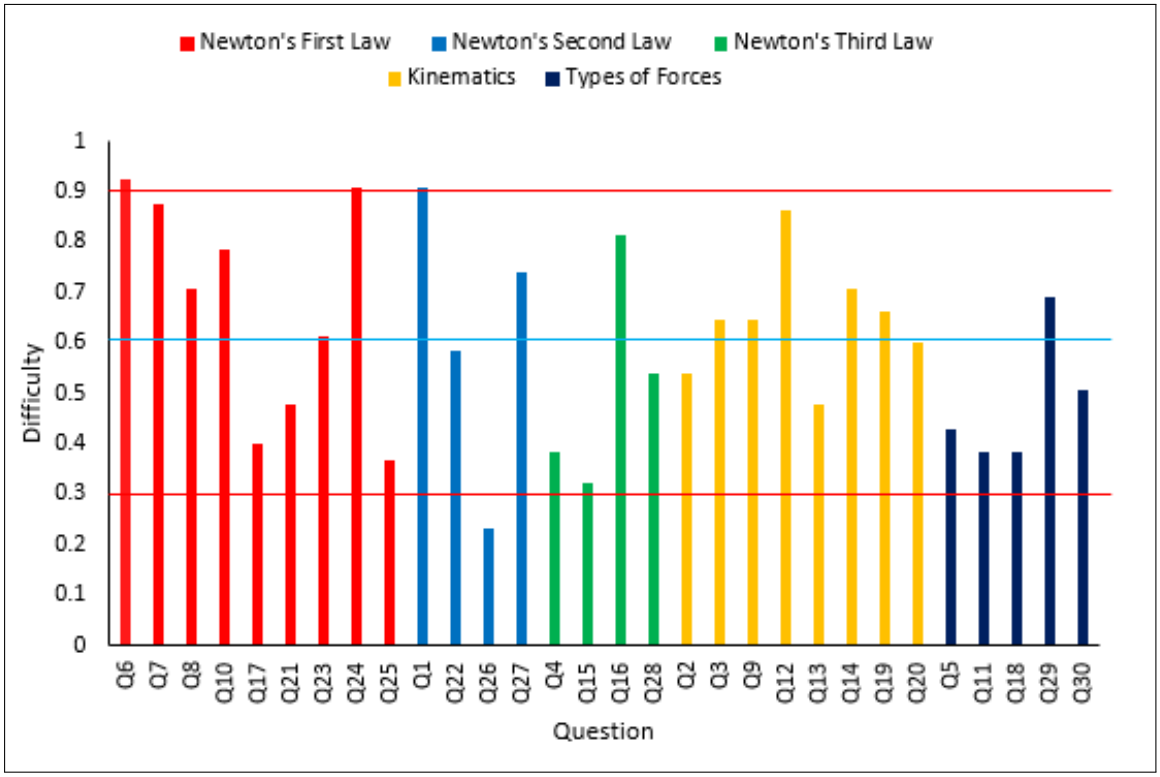


Figure 3.2: Graph showing the difficulty of the FCI questions from the responses given by the 2016-2017 pre-test cohort. The bars are coloured and grouped together based on the concepts tested by the questions. The red horizontal lines indicate the lower and upper bounds of the acceptable range of values for the difficulty (Ding and Beichner, 2009), and the blue horizontal bar indicates the mean value of the difficulty of all the questions.

From the 2016-2017 pre-test responses, the easiest question was Q6, which tests the concept of Newton's First Law. The other questions that tested the concept of Newton's First Law are Q7, Q8, Q10, Q17, Q21, Q23, Q24 and Q25 (shown by the red bars in Figure 3.2). The majority of these questions had difficulties that were above 0.5 (which is the preferred value of difficulty), which indicated that the cohort did not struggle with the questions testing understanding of Newton's First Law on the whole.

The hardest question was Q26, and this question tests the concept of Newton's Second Law. The other questions on the FCI which test the concept of Newton's Second Law are Q1, Q22 and Q27 (shown by the light blue bars in Figure 3.2). The cohort found Q1 and Q27 to be easier, but did not find Q22 to be particularly easy or difficult. Each of the questions tests Newton's Second Law in a different context, and it is possible that this led to the variability in performance on these questions.

The concept of Newton's Third Law was tested in Q4, Q15, Q16 and Q28 (shown by the green bars in Figure 3.2). Q4 and Q15 had difficulties which were below 0.5, and Q15 had a particularly low value, which indicated that the cohort struggled with half of the questions which tested understanding of Newton's Third Law. The concept of Kinematics was tested in Q2, Q3, Q9, Q12, Q13, Q14, Q19 and Q20 (shown by the orange bars in Figure 3.2). For the majority of these questions, the difficulty values were above 0.5, which indicated that the cohort did not struggle overall with the concept of Kinematics. The concept of Types of Forces was tested in Q5, Q11, Q18, Q29 and Q30 (shown by the dark blue bars in Figure 3.2). Most of these questions had difficulty values that were less than 0.5, which indicated that the cohort struggled to answer questions based on the Types of Forces concept.

The above findings showed that the 2016-2017 pre-test cohort had varying levels of performance when answering FCI questions testing different concepts prior to instruction. For the cohort, *Newton's First Law* and *Kinematics* were the easier concepts to answer questions on, whereas *Newton's Third Law* and *Types of Forces* were the harder concepts to answer questions on. This is consistent with findings from the literature, where Newton's Third Law is frequently reported to cause difficulties for students (Stocklmayer et al., 2012; Hughes, 2002; Brown, 1989; Yeo and Zadnik, 2000; Wells et al., 2019). The cohort had mixed success at answering questions based on the concept of *Newton's Second Law*. However, there was variation in difficulty between questions that were testing the same concept, and this indicated that other factors affected student performance on the questions. This point was previously raised by Stoen et al. (2020), who found that a variety of factors such as students' reasoning and problem-solving skills contributed to overall FCI score. As a final point, the mean difficulty of all the questions was 0.60, and this is represented by the blue horizontal line in Figure 3.2. This mean difficulty is close to the preferred difficulty value of 0.5, and this indicates that the FCI overall was at a suitable level of difficulty for the 2016-2017 pre-test cohort.

3.3.2 Findings from the 2016-2017 FCI post-test

The frequency of each of the different scores for the $N = 28$ completed post-tests were shown by the orange bars in Figure 3.1 in **Subsection 3.3.1**. Note that the frequency of each of the different pre-test scores (given by the blue bars) were also shown in Figure 3.1 for comparison. Unlike the 2016-2017 pre-test graph, the 2016-2017 post-test graph appears to be left skewed. This may be expected, since the cohort had studied the corresponding Newtonian mechanics content before attempting the FCI this time. The mean score was 21.89, and the median score was 23. The mean and median scores were both larger than the expected score from guessing of 6. In addition, the mean and median scores on the 2016-2017 post-test were higher than their corresponding values from the 2016-2017 pre-test, which is consistent with the reasoning given above for the graph's left skew.

As was the case for the 2016-2017 pre-test data, it is useful to look at student performance on the individual items on the test. To do this, the difficulty of each of the questions was calculated for the 2016-2017 post-test, and the results are shown in Figure 3.3 below, along with the concept that each question tests understanding of. The typology used to classify the post-test questions was the same as that used for the pre-test questions.

From the 2016-2017 responses to the post-test, the easiest question was Q7, which every student managed to get correct. In contrast, the hardest question for the cohort was Q17. Both Q7 and Q17 test the concept of Newton's First Law, and the other questions that tested this concept were Q6, Q8, Q10, Q21, Q23, Q24 and Q25 (shown by the red bars in Figure 3.3). The cohort generally answered these questions well, which meant that the easiest and hardest question for the post-test cohort corresponded to a concept that the post-test cohort found easier overall.

The other concepts are considered next. Newton's Second Law was tested in questions Q1, Q22, Q26 and Q27 (shown by the light blue bars in Figure 3.3) and students generally answered these questions well. The concept of Newton's Third Law was tested in Q4, Q15, Q16 and Q28 (shown by the green bars in Figure 3.3). There was a mix of easier and more difficult questions here, so the cohort did not appear to find Newton's Third Law particularly easy or difficult. The concept of Kinematics was tested in Q2, Q3, Q9, Q12, Q13, Q14, Q19 and Q20 (shown by the orange bars in Figure 3.3). All of these questions had difficulties that were above 0.5, which meant that

the cohort answered questions based on the concept of Kinematics well. The concept of Types of Forces was tested in Q5, Q11, Q18, Q29 and Q30 (shown by the dark blue bars in Figure 3.3). These questions all had difficulties values above 0.5, so the cohort did not struggle when answering questions based on the Types of Forces concept.

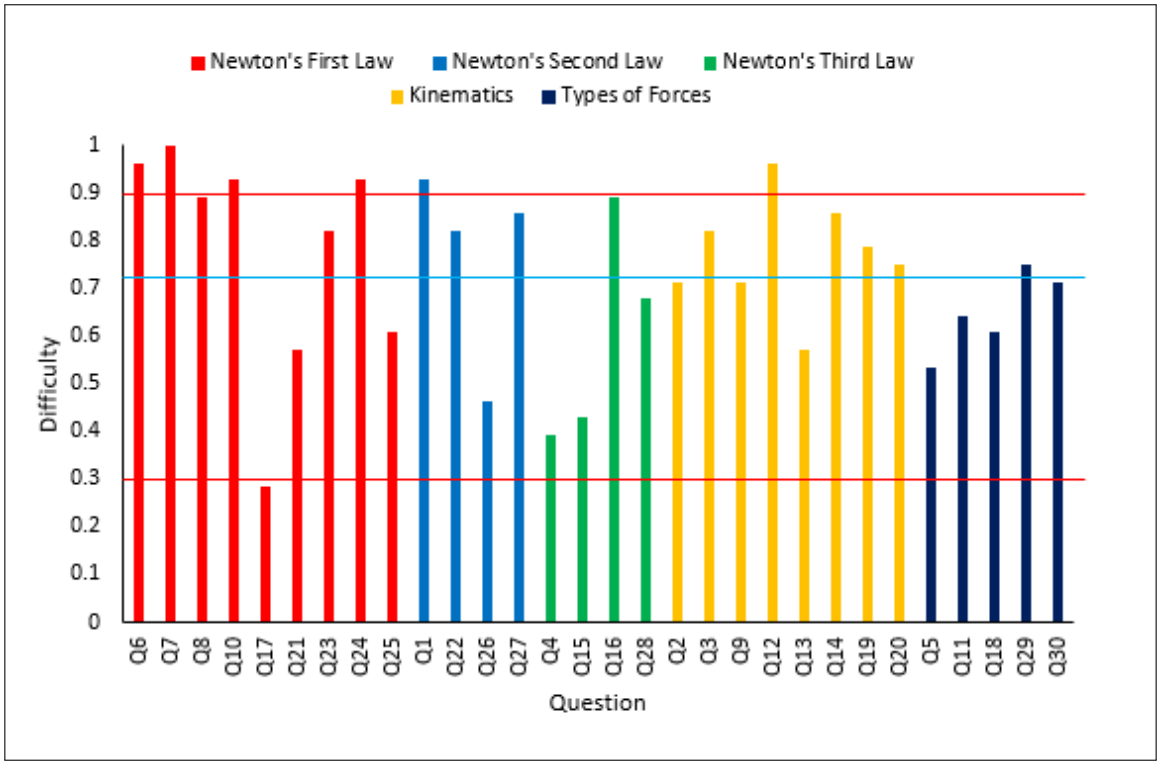


Figure 3.3: Graph showing the difficulty of the FCI questions from the responses given by the 2016-2017 post-test cohort. The bars are coloured and grouped together based on the concepts tested by the questions. The red horizontal lines indicate the lower and upper bounds of the acceptable range of values for the difficulty (Ding and Beichner, 2009), and the blue horizontal bar indicates the mean value of the difficulty of all the questions.

To summarize, the 2016-2017 post-test cohort managed to answer questions based on the concepts of Newton's First Law, Newton's Second Law, Kinematics and Types of Forces well. The cohort had mixed success at answering questions based on the concept of Newton's Third Law, meaning that there was still scope for the post-test cohort to improve their understanding of this concept. Comparing with the pre-test, the cohort continued to answer questions based on Newton's First Law and Kinematics well, and improved in their answering of questions based on the concepts of Newton's Second Law, Newton's Third Law and Types of Forces. In addition, the mean difficulty over all the questions was higher for the post-test cohort than the pre-test, which indicated that the post-test cohort scored higher on the FCI overall than the pre-test cohort.

This was reflected in the above findings, as the post-test cohort struggled with fewer topics overall than the pre-test cohort.

The mean difficulty of all the questions for the 2016-2017 post-test responses was 0.73, and this is represented by the blue horizontal line in Figure 3.3. This value is higher than both the preferred value for difficulty of 0.5 and the corresponding mean difficulty value for the 2016-2017 pre-test, which indicates that the FCI was an easier activity for the 2016-2017 post-test cohort. Findings from the post-test cohort are further compared and contrasted to the findings from the pre-test cohort in what follows.

Further comparison of the 2016-2017 pre-test and post-test findings

The difficulty values for individual questions can be compared between the pre-test and post-test cohorts. The post-test difficulty values were higher than the pre-test difficulty values for 29 out of the 30 FCI questions, meaning that these individual questions were easier for the post-test cohort. For Q17, the opposite trend was observed. This means that in the case of Q17, the post-test difficulty value was lower than the pre-test difficulty value, indicating that the post-test cohort found this question to be harder than the pre-test cohort. Q17 asks test-takers to compare the forces acting on an elevator as it moves up a frictionless elevator shaft, and it tests the concept of Newton's First Law. The post-test cohort did not struggle with other questions testing the concept of Newton's First Law, making the difficulties in answering Q17 an isolated case. As a result, it is possible that the low difficulty value on the Q17 post-test can be attributed to random fluctuations in the data.

The easiest question differed between the pre-test cohort and the post-test cohort, although the questions tested the same concept. For the pre-test cohort, the easiest question was Q6. This question requires test-takers to identify the path taken by a marble after it has been fired out of a frictionless channel onto a frictionless tabletop, and tests the concept of Newton's First Law. In addition, the post-test cohort also answered Q6 well. The easiest question for the post-test cohort was Q7, which asks test-takers to identify the trajectory of a steel ball after it has been thrown as a hammer, and tests the concept of Newton's First Law. Furthermore, the pre-test cohort also found Q7 to be a relatively easy question to answer. The pre-test and post-test cohorts both generally answered questions based on the concept of Newton's First Law well, so the easiness of Q6 and Q7 for both cohorts is an expected outcome.

The hardest question on the FCI was different for the pre-test and the post-test cohorts. For the pre-test cohort, the hardest question was Q26. This question asks test-takers to identify what happens to the speed of a box after the force applied to it is doubled, and it tests the concept of Newton's Second Law. As mentioned previously, the pre-test cohort had mixed levels of success at answering questions based on the concept of Newton's Second Law; it is hence possible that the situation presented in Q26 was difficult for the pre-test cohort to interpret. The post-test cohort did not have problems answering Q26, which is an expected outcome since the post-test cohort generally answered questions based on the concept of Newton's Second Law well. The hardest question on the FCI for the post-test cohort was Q17, and possible reasons for this question being particularly difficult for the post-test cohort were discussed previously. The pre-test cohort did not have problems with answering Q17, and this outcome is not surprising because the pre-test cohort answered most of the questions based on the concept of Newton's First Law well.

The similarity in performance on Q6 and Q7 between the pre-test and the post-test cohorts is an example of the two cohorts exhibiting similar behaviour, whereas the differences in performance on Q17 and Q26 give examples of the two cohorts behaving differently. Each of these cases provided useful information about student understanding, and this backs up the point made by Scott and Schumayer (2017) that correct and incorrect answers to FCI questions are both capable of providing a wealth of useful data to physics educators.

The findings discussed above indicated that the 2016-2017 *S217* cohort performed better on the FCI post-test than the pre-test. This is expected behaviour, since students should improve their performance after instruction, rather than degrade it. However, the two data sets compared for pre-test and post-test contained different numbers of students. In addition, confounding variables and selection effects such as only keen students doing the post-test and getting higher scores also affected these results. This means that the improvement observed in performance between pre-test and post-test could not be attributed to the instruction alone, although it was a contributing factor. Calculation of the normalized gain g and the normalized change c statistics provided a means of quantifying the effectiveness of the instruction, and this is the focus in the following subsection.

3.3.3 Findings from the 2016-2017 normalized gain and normalized change calculations

When calculating the normalized gain and normalized change, data were only used from the 27 students who completed both the pre-test and the post-test. The graph showing the pre-test and post-test scores of each of these 27 students is given in Figure 3.4 below.

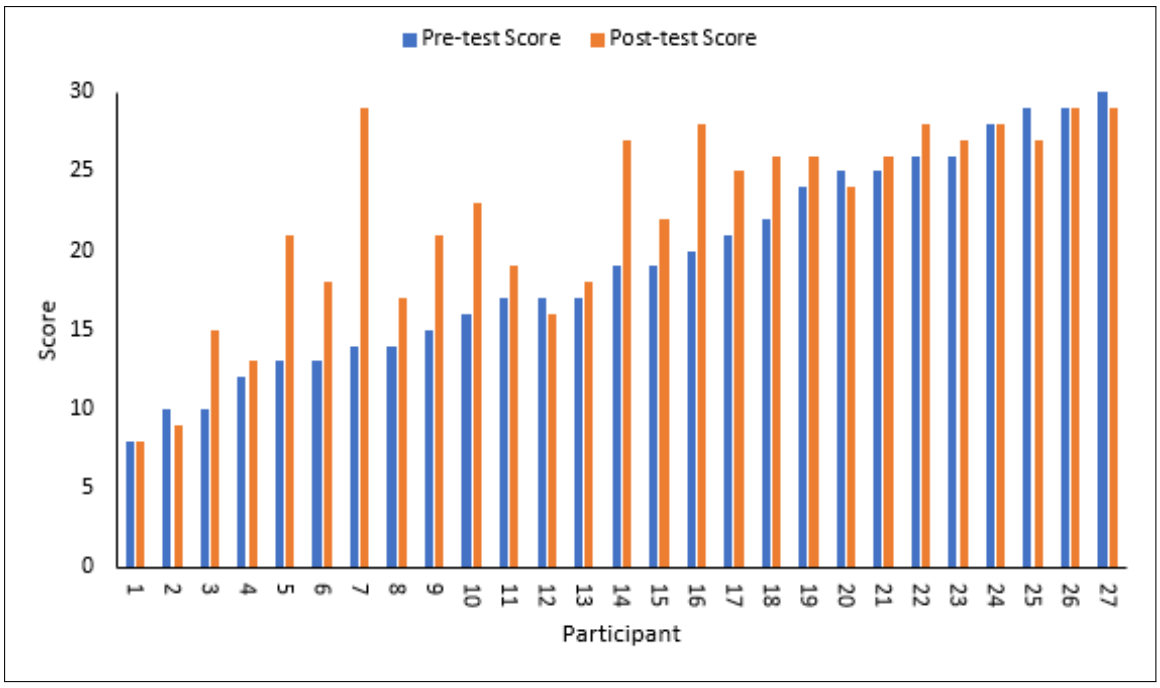


Figure 3.4: Graph showing the 2016-2017 FCI pre-test and post-test scores for each of the participants who completed both tests. The bars are colour coded based on whether they indicate the pre-test or the post-test score. The scores are presented in ascending order by pre-test score.

In Figure 3.4, the mean of the pre-test scores was 19.22, and the mean of the post-test scores was 22.19. These values are higher than the corresponding mean scores for the pre-test with $N = 65$ and the post-test with $N = 28$. In addition, there were no scores of zero on either the pre-test or the post-test. The results of the normalized gain and normalized change calculations using these data can be found in Table 3.1 below. The mean of the normalized gains and normalized changes were both positive for the 2016-2017 cohort, which indicated that the students did better on average on each of the individual questions in the post-test than on the pre-test; this backs up the previous finding that the mean question difficulty was higher for the post-test than the pre-test. In addition, the normalized gain and normalized change of the mean scores

were also both positive for the 2016-2017 cohort, which indicated that the students did better on the entire test after studying the material than before they studied the material; this agrees with the previous finding that the mean total score was higher for the post-test than the pre-test.

Statistic	Value
Mean of the normalized gains	0.19
Normalized gain of the means	0.27
Mean of the normalized changes	0.26
Normalized change of the means	0.27

Table 3.1: Normalized gain and normalized change values calculated using responses from the 27 students who completed both the 2016-2017 FCI pre-test and post-test.

The calculated values for the normalized gain and normalized change were positive, which implied that the cohort had improved their conceptual understanding of Newtonian mechanics after studying the corresponding *S217* course material. However, each of the statistics presented in Table 3.1 had values that were less than 0.3, meaning that the learning gain corresponded to the *low gain* band (Hake, 1998) for the 2016-2017 cohort; this indicated that further learning could have taken place. There are different ways to interpret this finding. In one interpretation, the *S217* teaching materials are held responsible for the low gain values, which implies that they could be modified to improve student learning of Newtonian mechanics. In a contrasting interpretation, properties of the data set used in the calculations are responsible for the low gain values. This can be explained as follows.

Because the *S217* FCI activities were optional, only the most enthusiastic students were likely to have done both the pre-test and the post-test, and these students are often also the most capable. As previously alluded to, this caused the pre-test scores used in the calculation to be higher than expected, and this made it difficult for the corresponding post-test scores to be significantly higher for these students. As a result, the learning gain and learning change calculations were not be able to yield *medium gain* or *high gain* results, since the students used in the calculations had less gain to get in the first place.

3.4 Conclusions

The study illustrated how a concept inventory can be used in a pre-test, post-test format to investigate the effectiveness of teaching methods. By administering the FCI digitally as a pre-test and a post-test survey in the academic year 2016-2017, performance data were gathered and analyzed. It was found that students generally improved in their performance after studying the material, although there are likely to be other confounding factors contributing to this trend. In addition, students were found to struggle with questions covering the concept of Newton's Third Law even after instruction, which is consistent with findings from the literature. By focusing attention to those areas where students struggled, it may be possible to develop even more effective teaching material for the *S217* module in the future. In the wider context of the current work, findings such as these form the basis of how a concept inventory can be used to test and develop new teaching methods.

3.5 Summary and looking ahead

Chapter 3 presented a case study of a multiple-choice concept inventory being used in practice. The findings were that students improved their performance on the FCI after studying the relevant material, but they still struggled with the concept of Newton's Third Law.

Chapter 4 outlines the methodology used to develop the *Alternative Mechanics Survey* (AMS), which is an adapted version of the FCI that makes use of free-response questions.

4 Development of the free-response Alternative Mechanics Survey

4.1 Rationale

The previous chapter examined how the traditional multiple-choice *Force Concept Inventory* can be used to find out which Newtonian mechanics topics students find difficult, and to assess learning gain. The overall aim of the research is to investigate whether automatically-marked free-response questions can be used to gain a better understanding of students' line of reasoning when answering physics concept inventory questions. To this end, a version of the FCI that makes use of free-response questions was developed. This instrument was dubbed the *Alternative Mechanics Survey*, in order to avoid confusion with the traditional multiple-choice FCI. The aim of the current chapter **(Chapter 4)** is to outline how the work presented in **Chapters 5, 6, 7 and 8** fits together to complete the iterative development process of the AMS.

4.2 Moodle Pattern Match

The *Pattern Match* question type of the Moodle question engine was used throughout the study. Free-response questions are written in Pattern Match by specifying the question wording and corresponding marking rules. An example of a Newtonian mechanics Pattern Match question is show in in Figure 4.1 below. The question is an example question based on the topic of *balanced forces*; it is not taken from either the FCI or the AMS.

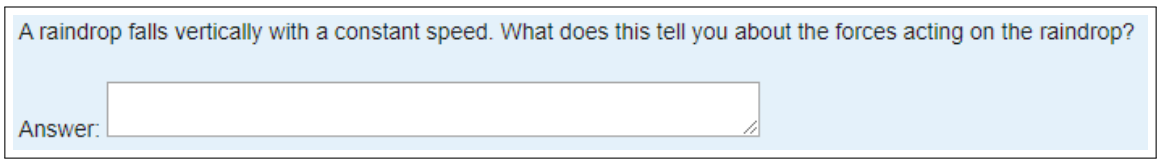


Figure 4.1: An example of a Pattern Match question.

Pattern Match marks the responses to free-response questions as follows. The question author specifies a set of marking rules, and each rule is designed to match answers to strings of characters (for example in words), with capability to also take account of other factors such as omitted and incorrectly ordered letters, truncation, word order and spacing. Some of these rules match to *positive* conditions that correspond to *correct* answers, whereas other rules match to *negative* conditions that correspond to *incorrect* answers. These rules are arranged from top to bottom by priority of

importance, with the highest priority rule at the top of the list. When an answer is marked by the system, it is matched against each rule in sequence until it either finds a match with one of the rules or does not match with any. When a match with a rule is found, the answer is marked based on whether the rule corresponds to a correct or an incorrect marking condition. A mark of 1 is awarded to an answer that matches with a correct marking rule; a mark of 0 is awarded to an answer that either matches with an incorrect marking rule, or does not match to any marking rule. Rules can also be used to generate appropriate feedback, and to give partial credit, although this function was not used in the study; instead, all questions were simply marked as either correct or incorrect.

Students can give answers in a variety of different ways, and Pattern Match has several features to accommodate for these possibilities. The question author can decide whether the answers need to be case sensitive, and also whether to permit the use of subscripts and superscripts in the answer. The question author can choose whether answers are required to be written in fewer than a certain number of words and whether or not to accept misspellings. Misspellings can be handled in two ways, both by the presence of a dictionary which checks whether it recognizes the words used in an answer, and by allowing reversed and omitted letters in words. In addition, Pattern Match gives the option for the question author to add their own words, which could be specialized vocabulary or synonyms not known to the system. There is a further option to convert certain characters into blank space; this facilitates the entry and marking of answers, as students may choose to include characters and symbols in their answers that the engine is not designed to deal with. Finally, there is an option to give a model correct answer to the question, allowing others editing the question to have a rough idea of the sorts of answers that should be marked as correct.

Automated marking schemes are more effective when student responses are used in their development (Jordan and Mitchell, 2009). The process of using student responses to develop Pattern Match marking rules is as follows. To start, the question author writes some basic computer marking rules for the question, based on their own experience and insight. In addition to this, the author writes a mark scheme which human markers can use to mark the question. Student responses to the question are then gathered, and these are marked by several human markers and by the computer marking rules.

A **master mark scheme** for the question is then developed by examining how the *majority* of the human markers decided to mark the question, with an expert marker acting as the arbitrator in the event of a tie. The master mark scheme is developed in this way because human markers are capable of being inconsistent and making mistakes. Furthermore, human markers inevitably interpret mark schemes in different ways (Butcher and Jordan, 2010). This means that a single human marker (including an expert) is not necessarily correct in their marking all of the time. The computer then compares the marks awarded by the master mark scheme to those awarded by the automated mark scheme, and gives a corresponding *marking agreement* as a percentage based upon the agreement between the two sets of marks. This percentage gives an idea of how well the computer marking rules are functioning, with a higher value indicating a better level of effectiveness.

The marked responses can be used to improve the computer marking rules. This is done by comparing the master scheme and computer marking of each individual response, and highlighting instances where the two markers disagree. There are two types of disagreement case. The first case occurs when the master scheme marks a response as correct while the computer marks the same response as incorrect, which is known as a *false negative*. Conversely, the second case occurs when the master scheme marks a response as incorrect while the computer marks the same response as correct, which is known as a *false positive*. Modifying the marking rules to remove false negatives and false positives makes the computer marking align better with the master mark scheme, and this increases the effectiveness of the computer marking rules. This is achieved by adding synonyms and other correct cases to account for the false negatives, and by adding negations to cancel the false positives.

Changes made to the marking rules using the above approach are based on cases of false negatives and false positives found in the marking of a specific set of responses. It is hence possible that the modified marking rules work well for the set of responses that they were designed from, but not for other sets of responses. In the literature, this is known as an *over-fitting* problem (Zehner et al., 2016). As a result, the modified marking rules need to be tested against a new set of responses to check if they are effective. This starts the process of comparing human and computer marking again, and the marking rules may need to be modified further if false negative and false positive cases arise from marking the new set of responses. Marking rule development

is thus an iterative process, with more responses being required to repeat the process if the marking agreement is not high enough.

The required level of marking agreement is dependent upon the purpose of the Pattern Match questions. A benchmark of 95% marking agreement is indicative that the marking rules are able to mark on the same level as an expert human, or better (Jordan, 2012b), making it a logical cut-off to aim for when authoring marking rules. To reach this level, several iterations of the above process are often needed, with typically hundreds of responses being required to get the marking agreement to a high enough level (Jordan and Mitchell, 2009). In addition, the above process highlights the difference between the initial step and subsequent steps of developing the marking rules. In the initial step, the marker comes up with a few rules using experience and understanding of the subject, whereas in the iterative steps the marker uses student responses to improve and test the marking scheme.

4.3 The Alternative Mechanics Survey development process

Previous work done by Doctor Ross Galloway at the University of Edinburgh; Doctor David Sands at the University of Hull; and Doctor Christine Leach and Professor Sally Jordan at The Open University investigated whether the 30 questions on the traditional FCI could be asked in free-response format. To do this, the group adapted all the FCI questions into free-response format, making a deliberate decision to keep the wording as close as possible to that used in the original multiple-choice format. In addition, three questions (AMS Q3, Q7 and Q19 in Tables A.1 and A.2 of **Appendix A**) were added to test concepts linked to those in adjacent questions. Galloway and Sands collected written responses to these 33 questions from a total of 326 students in their introductory physics courses. Leach took these responses and authored corresponding automated marking rules using the Pattern Match technology of the Moodle question engine, finding that some questions were unsuitable for Pattern Match. This initial work was preliminary, and it provided a *proof of principle* for the development of a version of the FCI which makes use of free-response questions.

The author of this thesis took this idea up by developing the *Alternative Mechanics Survey* (AMS). The author considered all 33 questions from the preliminary work for inclusion in the AMS, and selected to keep 18 of the questions in free-response format, as Leach had previously demonstrated that these questions were suitable for

direct adaption into free-response format in the preliminary work. As a result, the author used the corresponding responses to write Pattern Match marking rules for these questions. This was a deviation from the general process outlined in **Section 4.2**, and this approach was taken because responses to the questions had already been collected. An example of one of the questions is shown in Figure 4.2 below.

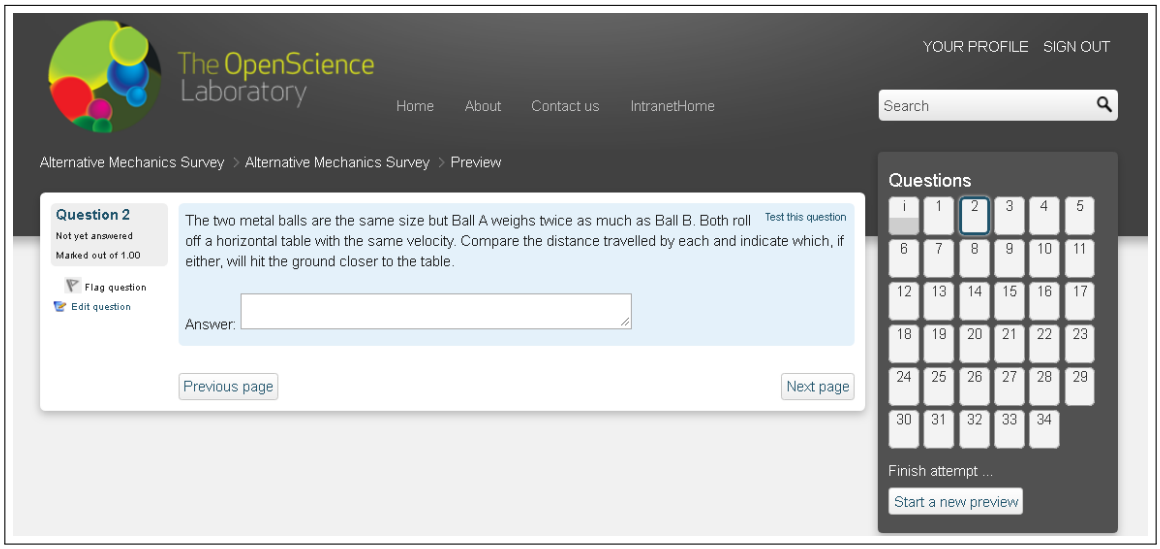


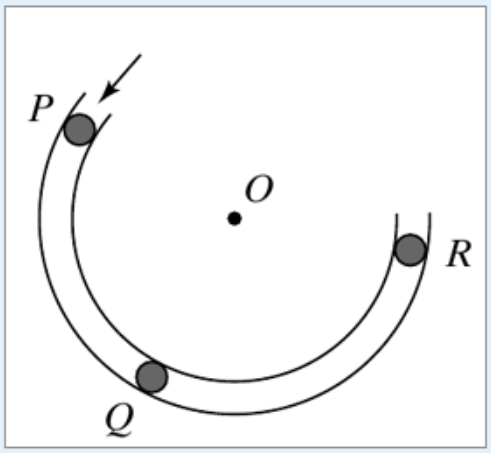
Figure 4.2: Question 2 of the AMS, which is a free-response question. This question builds on a scenario established in Question 1 of the AMS, which is based on Q2 of the FCI.

In addition to the 18 free-response questions, the author reverted 11 of the questions to their original multiple-choice FCI form. Seven of these 11 questions were based on trajectories, which are difficult to describe in words. For the other four of these 11 questions, the correct and incorrect responses had been found difficult to disentangle using Pattern Match syntax, making the authoring of effective marking rules difficult. An example of one such question was Q15 of the AMS, which asks test-takers to identify the force (or forces) acting on a ball after it has been thrown into the air. The sought correct answer to the question was *weight* (or equivalent), but students frequently gave answers which incorrectly identified *weight* and *gravity* as separate forces, or made reference to additional forces acting on the ball after it had landed. These answers contained a mixture of correct and incorrect information, a known problem for the automatic marking of free-response questions (Mitchell et al., 2002).

The author converted the remaining 4 out of the 33 questions into the **multiple-response question** format, which allow multiple options to be selected. Each of these questions asked students to identify the forces acting in a given situation, and

these were difficult to write marking rules for because of the wide variety of responses given to them. These questions were turned into multiple-response questions rather than reverted to their original multiple-choice FCI forms; this is because students still need to identify the forces for themselves when answering the multiple-response version of the question. Unlike multiple-choice questions, multiple-response questions cannot be answered using *eliminate and guess* strategies, which means that they are able to provide more information about students' understandings and misconceptions when answered. For comparison, the multiple-response version of a question from the AMS and the multiple-choice version of the corresponding question from the FCI are shown in Figures 4.3 and 4.4 below.

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high speed into the channel at *P* and exits at *R*.



Which of the following forces are acting on the ball when it is in the frictionless channel at point *Q*?

Select one or more:

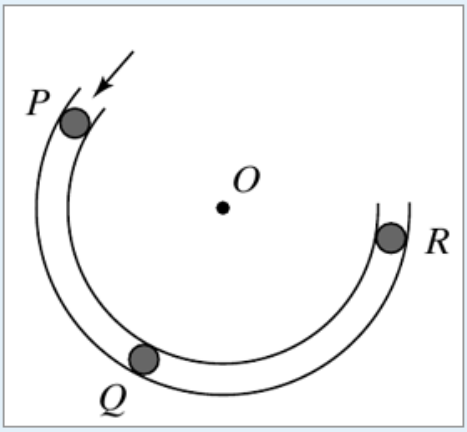
- ☐ A downward force of gravity.
- ☐ A force pointing from Q to O.
- ☐ A force in the direction of motion.
- ☐ A force pointing from O to *Q*.
- ☐ An upward force from the table.

Figure 4.3: A multiple-response question from the AMS. The multiple-choice version of this question from the FCI can be found in Figure 4.4.

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with center at *O*. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. Forces exerted by the air are negligible. A ball is shot at high speed into the channel at *P* and exits at *R*.

Consider the following distinct forces:

1. A downward force of gravity
2. A force exerted by the channel pointing from *Q* to *O*.
3. A force in the direction of motion.
4. A force pointing from *O* to *Q*.



Which of the above forces is (are) acting on the ball when it is within the frictionless channel at position *Q*?

Select one:

- A. 1 only.
- B. 1 and 2.
- C. 1 and 3.
- D. 1, 2, and 3.
- E. 1, 3, and 4.

Figure 4.4: A standard multiple-choice question from the FCI. The multiple-response version of this question from the AMS can be found in Figure 4.3.

Version 1 of the AMS was assembled by putting these questions into a 33 question test using Moodle. In addition to these questions, the test included an information screen and an open-ended question allowing participants to give feedback about the test. Version 1 needed to be tested for *validity*, to find out whether it was capable of doing what it was designed to do. In the context of the current work, this meant investigating qualitatively whether the FCI questions could be asked in the free-response format. To conduct this investigation, formal usability testing with corresponding interviews was carried out with eight students, to find out whether they reacted adversely to being asked the FCI questions in a free-response format. Details of this usability testing can be found in **Chapter 5**.

Version 1 also needed to be tested for *reliability*, to find out whether it was capable of producing the same results when used over and over again. Since the AMS is made up of questions and marking rules, both of these needed to be tested for reliability. To obtain the data to conduct this testing, responses were gathered through the Open-

Science Laboratory (OSL) by giving Version 1 of the AMS to a total of 328 high-school students and undergraduate university students. For the reliability testing itself, the questions were tested using the *Classical Test Theory* (CTT) approach, whereas the marking rules were tested using the *Inter-Rater Reliability* (IRR) approach. Details of these CTT and IRR calculations can be found in **Chapter 6**.

Findings from the usability testing and the CTT study were used to make changes to the Version 1 questions, and findings from the the IRR study were used to make changes to the Version 1 marking rules; taken together, this iterated Version 1 into Version 2 of the AMS. To further the development of the AMS with respect to the design priority of using free-response questions, it was decided that every question in Version 2 would require students to enter an answer into a free-response box. For the seven trajectory-based multiple-choice questions from Version 1, the multiple-choice options A, B, C, D and E corresponding to the different trajectories are still given, and the student only needs to enter one of these letters into the free-response box to answer. An example of such a question is shown in Figure 4.5 below, and this type of question is referred to as *free-response (letter)* throughout the rest of the thesis.

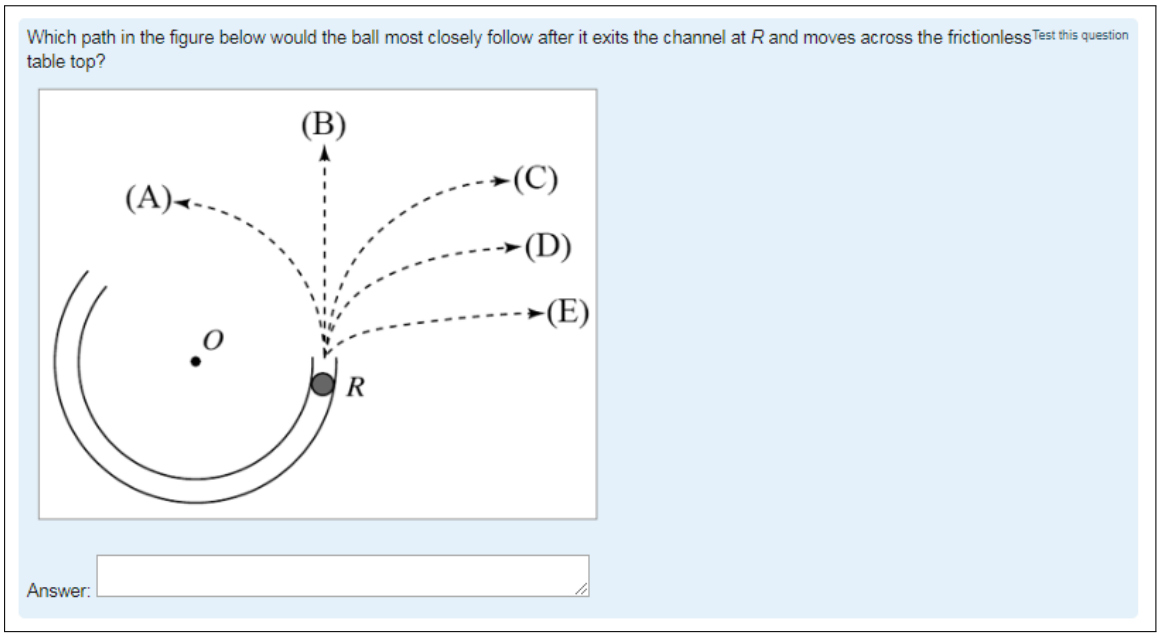


Figure 4.5: A free-response (letter) question from the AMS.

The four multiple-choice questions from Version 1 which were not trajectory based were altered to be standard free-response questions; these were given marking rules based on the marking rules of similar questions from the AMS. An example of one such question is given in Figure 4.6 below. Note the obvious difference in the amount of

detail that needs to be entered into the free-response box to answer the free-response (letter) questions and the standard free-response questions.

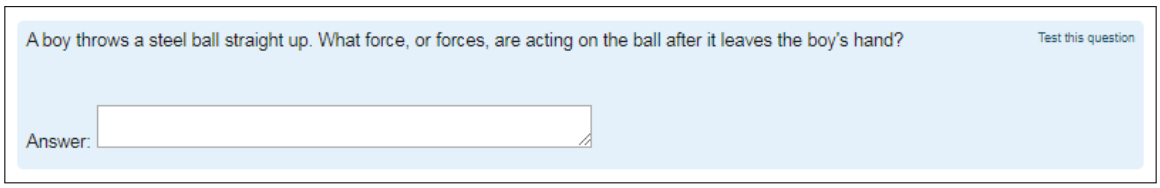


Figure 4.6: A standard free-response question from the AMS.

The usability testing conducted with Version 1 found it was a valid approach to ask the FCI questions in the free-response format; hence the validity testing did not need to be repeated for Version 2 of the AMS. However, since some questions had been changed from multiple-choice to free-response format between versions, the reliability established by the CTT testing on Version 1 could not be assumed to carry over to Version 2 *a priori*; this meant that the CTT strand of the reliability testing needed to be repeated for Version 2. In addition, issues with some of the marking rules were highlighted by the IRR study carried out on Version 1, meaning that the IRR strand of the reliability testing also needed to be repeated for Version 2. To obtain the data to conduct this testing, responses were gathered through the OSL by giving Version 2 to a total of 81 high-school students and undergraduate university students. Findings from the CTT study were used to make changes to the Version 2 questions, and findings from the IRR study were used to make changes to the Version 2 marking rules. Taken together, this iterated Version 2 into Version 3 of the AMS. Further details of these CTT and IRR findings can be found in **Chapter 7**.

From the CTT study carried out on Version 2, it was found that the AMS questions were functioning at the desired level, so the CTT strand of the reliability testing did not need to be repeated for Version 3, as the question set was now stable. However, there were still issues with the marking rules highlighted by the IRR study conducted on Version 2, meaning that this strand of the reliability testing needed to be repeated for Version 3. To obtain the data for this study, responses were gathered through the University of Cambridge's **Isaac Physics** platform (Isaac Physics, 2020) by giving Version 3 questions to a total of 118 high school students; an example of an AMS question as it appeared when administered on the Isaac Physics platform can be found in Figure 4.7 below. Findings from this study were used to make changes to the Version 3 marking rules, which were in turn used to iterate Version 3 into Version 4 of

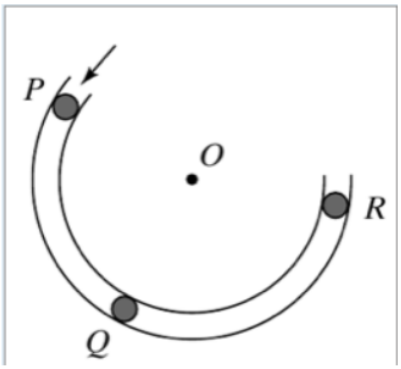
the AMS. More details of these IRR findings can be found in **Chapter 8**.

Part A

Part 1

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high speed into the channel at P and exits at R.

Figure 1



A circular frictionless channel

Answer Now

Identify the force or forces acting on the ball after it emerges from the track at R.

Its weight and the normal contact force from the table

Figure 4.7: An example of an AMS question as it appeared when administered on the Isaac Physics platform.

Version 4 of the AMS is the final one developed as part of the author's doctoral research and is one of the outputs of the project. A flowchart illustrating the process used to develop Version 1 into the final version is shown in Figure 4.8 below; a breakdown of how the questions on different versions of the AMS map to one another, and to the original FCI questions, is given in Tables A.1 and A.2 of **Appendix A**; and the final version of the AMS questions and marking rules can also be found in **Appendix A**. Note that the standardized AMS question numbering given in Tables A.1 and A.2 is used to refer to AMS questions throughout the remainder of the thesis.

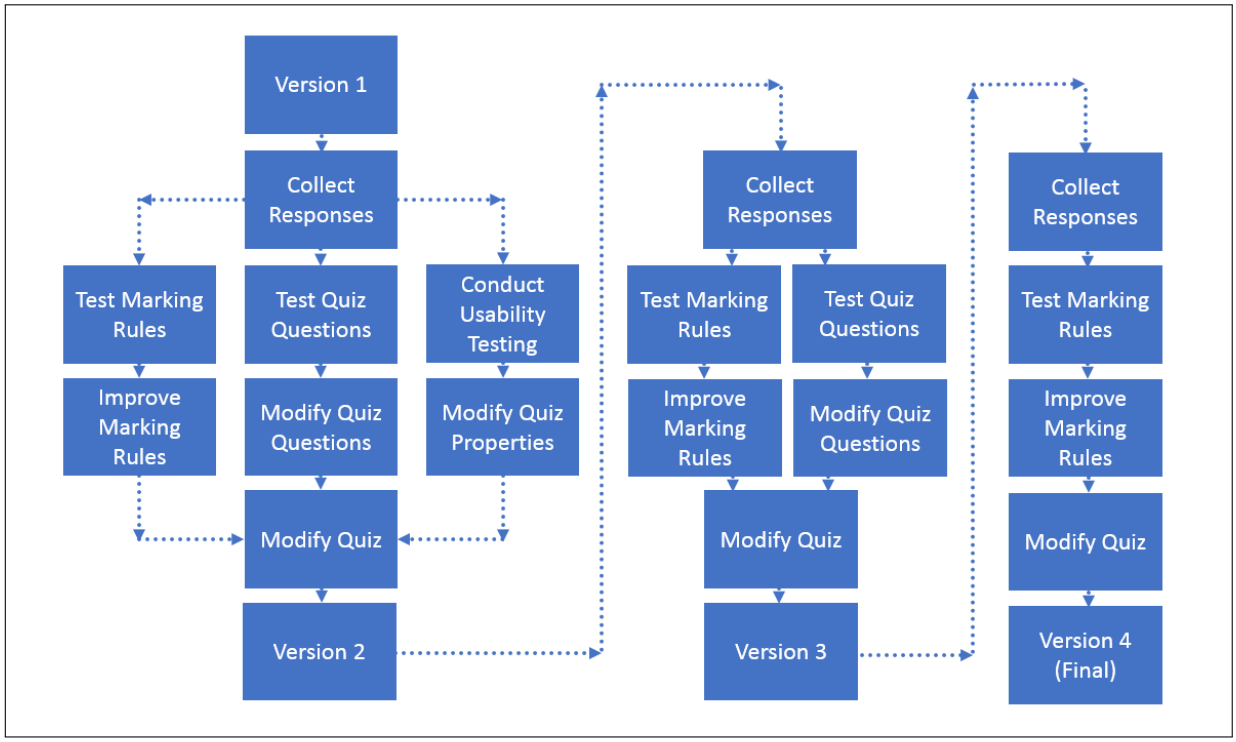


Figure 4.8: Flowchart showing the development process of the AMS.

4.4 Summary and looking ahead

Chapter 4 outlined the resources and process used to develop the AMS, as well as how the questions on the different versions of the AMS correspond to one another, and to the questions on the original FCI.

Chapter 5 focuses on the qualitative approaches used to test the AMS for validity. It presents findings from data gathered through the usability testing conducted with Version 1 of the AMS in the academic year 2017-2018.

5 Usability testing of the Alternative Mechanics Survey

5.1 Rationale

The questions on the original multiple-choice FCI have been widely validated but approximately half of these questions have been replaced by free-response counterparts in the AMS. The validity of these free-response questions (meaning that they measure what they are intended to) and consequently the entire AMS cannot be assumed to follow from the validity of the original FCI *a priori*. Since one of the key goals of this research is to develop concept inventories that are both accepted by the Physics Education Research community and widely used, it follows that a validity study must be carried out on the AMS. This can be done by investigating student reaction to the free-response questions, as this will gauge whether the free-response questions are able to do the same job as their multiple-choice counterparts.

When the FCI is administered in its regular format, students never find out how they have done on either the pre-test or the post-test because the concept inventory is instead designed to give the instructor feedback on the effectiveness of their teaching. Another of the key goals of the overall research is to develop new types of concept inventories, and these may serve wider educational purposes then the previously developed concept inventories. As a result, student reaction to being given feedback on the AMS was also investigated, as a way of seeing whether there was any scope to make use of the AMS as a teaching tool.

5.2 Methods

5.2.1 Data collection

The investigation took the form of a formal usability testing (Barnum, 2010). The methodology was chosen to suit the purpose of the study, and has previously been used to investigate student reaction to free-response questions (Jordan, 2012a). The study was conducted in the usability laboratory at the OU, which features a human computer interaction lab and a live observation room. The human computer interaction lab contains a computer on which participants can trial new software and study materials, as well as a webcam and audio link which allows the participants' actions and commentary to be both streamed and recorded. The live observation room allows viewers to watch the participant and their screen in real-time via the webcam and audio link.

As a precursor to the main usability testing study, the AMS was trialled with four subject experts (Norton, 2017; Croston, 2017; Mackintosh, 2017; Eden, 2018) in a think-aloud setting, and it was found that the AMS questions were interpreted in the desired way. After obtaining approval for the work from the University's *Student Research Project Panel* and the *Human Research Ethics Committee*, eight participants were selected for the study. The opportunity to be involved in the investigation was offered to students on the OU's second-year physics modules *S217 Physics: From Classical to Quantum* and *SXPA288 Practical Science: Physics and Astronomy*, as well as to first-year PhD students in the OU's School of Physical Sciences. An email message was sent to potential participants, and this gave details about the logistics of the usability testing, but did not give any details about the background of investigating use of free-response questions in concept inventories. Four undergraduate students and four postgraduate students were selected, being the first eight participants to respond. They were each given an *Amazon* voucher worth £20 in appreciation of their involvement.

The qualitative study thus made use of eight participants, and each testing session lasted about one hour. Testing and interviewing a greater number of participants would have been useful, but in order to maintain the project schedule, the usability tests were limited to these eight participants; this is acknowledged as a limitation of the study, as it does not allow for ready subdivision of qualitative data by specific characteristics of the participants (such as study experience or demographic). However, the method chosen to analyze the data, *Thematic Analysis* (see **Subsection 5.2.2**) is valid for small numbers of participants from whom rich data has been gathered, as in this case.

The trial participants had different levels of study experience and different subject backgrounds. For the undergraduate students, all four were studying the OU's second-year physics module *S217 Physics: From Classical to Quantum*, although they were working towards a variety of different physics-related qualifications. For the postgraduate students, one had studied geology as an undergraduate, one astronomy, one physics, and one chemistry. This meant that the participants had different levels of prior exposure to the test material, and with the small numbers involved, the results of such division would not be statistically significant. No conscious effort was made to gather participants of different ethnicities or genders, since investigating specific demographics was not the aim of the study. In what follows, the participants are identified as P1 to P8, with P1 to P4 being current OU undergraduate students, and P5 to

P8 being postgraduate students. Participant P7 was additionally a former OU undergraduate student. One of the participants was female, and the other seven participants were male.

Before taking the AMS in the OU usability laboratory, the participants were briefed on the structure of the usability testing. They were told that the semi-structured interview would be conducted after completion of the AMS, but they were not told any background about the FCI or concept inventories. The AMS was administered onscreen, and some instructions pertaining to the free-response questions were given in an information screen at the start of the AMS, as shown in Figure 5.1. This information explained that answers of a few words would be sufficient to answer the free-response questions, and that answers should be no longer than twenty words. Figure 5.2 shows the first question screen, and participants could receive feedback at the end of the AMS by clicking the *submit all and finish* button at the bottom of the final screen. The AMS questions given to the participants in the usability testing study can be found in **Appendix C**.

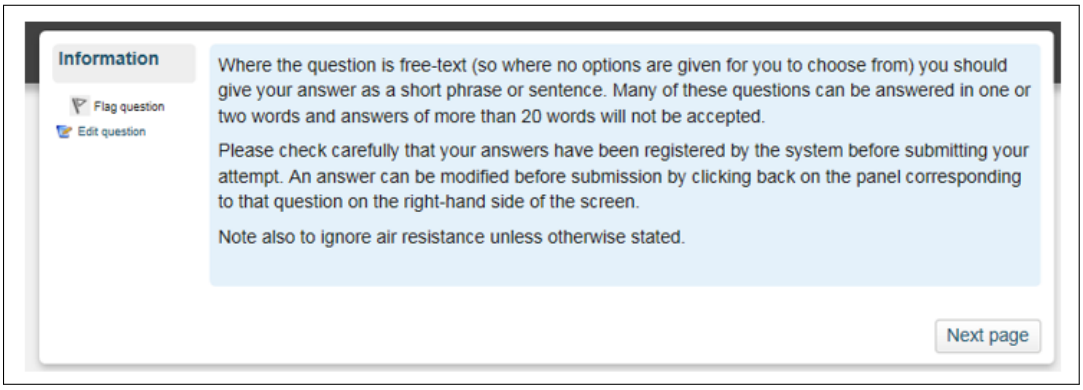


Figure 5.1: The instructions given at the start of the AMS.

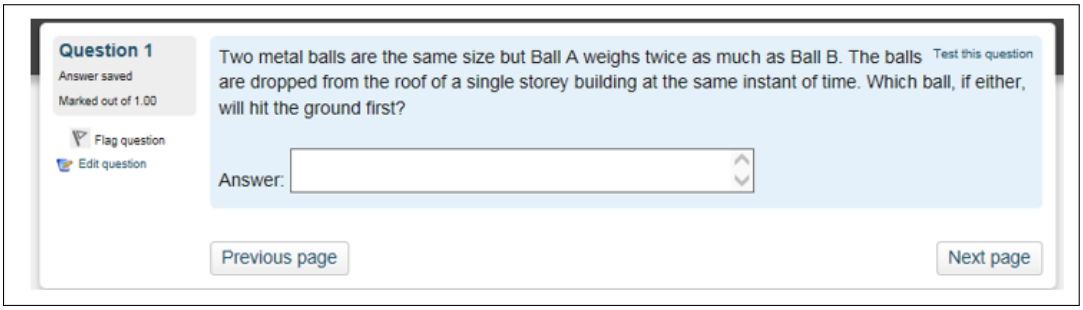


Figure 5.2: An example of a question on the AMS.

During the study, participants worked through the AMS while being recorded, and were watched remotely from the live observation room by two members of the research team. Participants were free to think aloud if they wished, but they were not given explicit instructions to do so, because this can affect the process being observed (Dockter and Mestre, 2014). After completion of the AMS, a semi-structured interview was conducted, with this interview also being recorded. No time limit was placed on the AMS, but participants were told that it would probably take between 30 minutes and 1 hour to complete.

Feedback was given to participants after they had completed the AMS and it was only provided on the version of the AMS used in this study; in line with common practice, no feedback is given on versions of the AMS provided to students more widely. The feedback provided was limited to knowledge of whether the answer was correct or incorrect, plus the correct answer for multiple-choice and multiple-response questions. It is relevant that the OU's distance-learning students usually get detailed feedback when they complete assessment tasks, and the majority of the participants were current or former OU students. It is hence likely that these participants were expecting some sort of feedback when they worked through the AMS.

In addition, in order to give direction to the semi-structured interviews that followed the usability observation of each student, some feedback was provided to the students on their AMS answers. This prevented the interviews from being derailed by participants wanting to check their answers, which allowed the focus to be placed on a discussion of whether the questions had been interpreted and interacted with in the intended way from a *usability testing* standpoint. Similar approaches have previously been used to add structure to interviews in the context of evaluating the effectiveness of remote teaching laboratories by Scanlon et al. (2004) and Nickerson et al. (2007).

There were three components to the data from these trials. The first were the answers given by the eight usability laboratory participants to the AMS questions; the second were the responses given by the eight usability laboratory participants to the interview questions; the third were responses to Q34 of the AMS gathered from the large-scale administration of Version 1 of the AMS. Q34 was a qualitative question added to the end of the AMS, asking how the students found the different question types on the AMS; a screenshot of the question can be found in Figure 5.3 below.

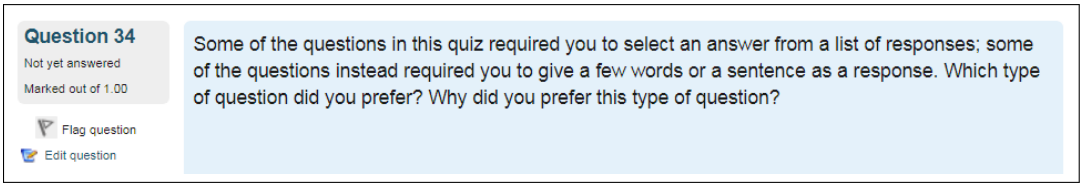


Figure 5.3: Q34 of the Version 1 of the AMS

The answers given by the participants as they worked through the AMS, the transcripts of the interviews, and the responses to Q34 were each rich qualitative data sets. *Thematic Analysis* (Braun and Clarke, 2006; Braun et al., 2014) was used on the interview transcripts to find underlying *themes*, with data from the AMS answers given used to support this, and the responses from a wider range of test-takers to Q34 were used for triangulation.

5.2.2 Thematic analysis

Thematic Analysis can be used when a small number of participants produce a rich qualitative data set. The aim of Thematic Analysis is to reduce this rich data set into an interpretable form, and these are the eponymous themes of the method. Themes emerge from the investigator's assimilation of the entire data set (here the answers to the AMS questions, interview transcripts, and film of the AMS sessions), which prevents arbitrary conclusions from being drawn. The interview questions asked to the participants in the AMS usability testing study can be found in **Appendix G**.

Thematic Analysis is a robust and widely used methodology which has been used in a range of contexts, including the analysis of forum posts (Smedley and Coulson, 2017), focus group data (Varagona and Hold, 2019; Garcia and Hamilton-Giachritsis, 2017), blog entries (Castro and Andrews, 2018), open-ended question responses (Grogan and Jayne, 2017), interview data (Robertson et al., 2018), and literature (Filia et al., 2018). Thematic Analysis has the capability to extract meaningful conclusions from a rich data set drawn from a small number of participants, and this makes it suitable for use in the current study. The version of Thematic Analysis outlined by the University of Auckland (2017) was used here, which has the following 6 steps:

• Step 1: Familiarization with the data. This step requires the data to be organized, transcribed and read through. In the study, this step was completed by transcribing the recordings of the participants' AMS attempts and interviews from the usability

lab.

• Step 2: Coding. Labels are assigned in the form of codes that identify key features of the data set. In the study, this step was completed by coding parts of the transcripts with keywords and phrases that summarized the content.

• Step 3: Search for themes. Within the identified codes, broader patterns of meaning must be found. The data must then be collected under identified themes. In the study, this step was completed by grouping together codes that were related, and discerning overlying themes for these code collections.

• Step 4: Reviewing themes. The themes need to be tested against the data set, and some themes may need to be refined or discarded. In the study, this step was completed by going back through the codes to see if anything had been missed or misplaced.

• Step 5: Defining and naming themes. A detailed analysis of each theme is conducted, and a name must be assigned to each theme. In the study, this step was completed by looking in-depth at the codes associated with each of the themes, and checking the evidence supporting each theme from the transcripts.

• Step 6: Writing up.

The steps are sequential, with each step building on the previous one. However, Thematic Analysis can be an iterative process, and different steps can be returned to as necessary throughout the analysis. The objective is to reduce the large qualitative data set into a set of results that can be analyzed.

It is of note that the author was the only one who went through the data and coded it, which is a limitation of the study. To counteract this, the codes were enumerated and gathered together under themes which exhibited broader patterns of behaviour. This approach was taken in order to mitigate the effect of annotator bias, and to prevent arbitrary conclusions from being drawn from the data. To make the meaning of the themes clear, the statement of each theme gives the main conclusion from the underlying codes associated with it, and this resulted in themes which were longer than usual. To facilitate with interpretation, a brief description of each theme is given when the corresponding findings are discussed in **Subsections 5.3.1 to 5.3.6**.

5.3 Results and Discussion

For the interview data, Thematic Analysis identified 17 codes, which grouped into 6 themes. These themes were *Use of free-response questions supported deep learning*, *Interpretation of the AMS instructions affected answer length*, *The idea of being marked by a computer did not affect answer structure*, *Participant reaction to the usability of the AMS was mostly positive*, *Reactions to the AMS depended upon what participants thought it was for*, and *Limited feedback was a useful addition to the AMS*. Each of these are discussed in **Subsections 5.3.1 to 5.3.6** below, with supporting quotes from the interviews and typed responses to the AMS questions referred to where appropriate. Additionally, the findings from the Q34 responses are triangulated with the interview data where this is relevant.

5.3.1 Findings related to the *Use of free-response questions supported deep learning* theme

There were 5 different codes associated with the *Use of free-response questions supported deep learning* theme, and this theme was coded 44 times overall. This theme covers participants' reactions to answering the free-response questions, and their comparisons of these with other question types. The codes associated with the theme are presented in Table 5.1. Note that unless otherwise stated, the occurrence of the codes was more or less equal between the eight participants. In addition, codes are assigned a label (such as C1) for ease of reference in the text. These conventions are applied in Tables 5.1, 5.5, 5.6, 5.7, 5.9 and 5.10.

Code	Number of times coded
Free-response questions made them think (C1)	7
Free-response questions encouraged them to be creative (C2)	6
Multiple-choice distractors can be misleading (C3)	9
Selected-response makes multiple-choice easier (C4)	12
Could see the potential benefits of both multiple-choice and free-response questions (C5)	10

Table 5.1: Codes associated with the *Use of free-response questions supported deep learning* theme.

From the codes C1 and C2, participants identified that free-response questions gave them the chance to express answers in their own words. As a result of this, answering the free-response questions was perceived as a creative process which required careful thought. In addition to this, participants P2 and P6 noted in the interview that they preferred the free-response questions. In the case of P2, this preference was because the participant believed that free-response questions help the test-taker to learn:

"Mostly it makes you think about the answer more, and it allows you to learn something yourself - Whereas, when you have some pre-prepared answers, like in multiple-choice, then there's only one possible answer, and you have a probability of getting it right - You may not necessarily learn as much" [P2].

In the case of P6, no reason was given for this preference.

With respect to code C3, participant P1 noted that the distractor options provided by the multiple-choice questions might draw students into giving an incorrect answer, which the they considered to be an unfair way of trying to make the questions harder:

"At least one of the questions was asked in a way where it was guiding you towards the wrong answer. So, I think that if the marker, well, if there's multi-choice for the answers then the question needs to be harder - So, I wouldn't say that it needs to be misleading" [P1].

In contrast to this, code C4 highlighted that participant P4 developed a strategy for answering the multiple-choice questions on the AMS; P4 would look at a multiple-choice option that seemed to be right, and then compare this answer to the other available options before making a final choice. It is of note that such a strategy could not be employed to answer a free-response question.

In the interviews, participants P5 and P7 identified that they preferred the multiple-choice questions. In the case of P5, this was because the multiple-choice questions provided the participant with guidance on how to answer if they were unsure:

"Yeah, it was definitely the multiple-choice types ones, simply because when I was unsure either on the question, or my reasoning, I could use the range of answers to kind of narrow down the possibilities" [P5].

For the case of P7, no reason was given for this preference.

In code C5, participants P1 and P4 saw the advantages of both question types. In the case of participant P1, they noted that multiple-choice questions were easier for the students, but that free-response questions provided more information for the markers:

"Yeah, well, multi-choice is always easier. And, its easier for the student, but the long descriptive answers are probably more useful for the marker, because they allow the student to demonstrate more understanding" [P1].

In the case of participant P4, they enjoyed having a mix of question types, and noted that free-response questions do not need to necessarily entirely replace the use of multiple-choice questions.

Findings from the Q34 responses: (a) Cases where students preferred free-response questions

Overall, $N = 229$ students who did Version 1 of the AMS gave responses to Q34. Within these responses, it was found that $N = 44$ preferred free-response questions. The reasons given by the students for this preference are given in Table 5.2 below.

Reason for preferring FRQs	Tally	Meaning
Allowed you to think/write own words	27	Refers to the student being able to think for themselves, and to write down their own answer to the question.
Tests understanding	8	Refers to the students identifying that free-response questions provide a better test of student understanding than the multiple-choice counterparts.
MCQ makes them doubt themselves	3	Refers to the idea that the options given by multiple-choice questions can cause the student to doubt that the answer that they have given is correct.
Quicker to answer	2	Refers to the idea that free-response questions are quicker to answer than their multiple-choice counterparts.
No reason given	4	Refers to instances where the student stated that they preferred free-response questions, but did not give any further explanation as to why.

Table 5.2: Reasons for students who did Version 1 of the AMS preferring the free-response question type.

Some students liked the fact that the free-response questions allowed them to think for themselves, and to write their own answers. In addition to this, some students preferred the free-response question type as they felt that it better tested their understanding of the material. Notably, these students' reasons for the preference of free-response questions were similar to the positive aspects of free-response questions identified in codes C1 and C2 by the participants in the usability laboratory study:

"Liked being able to enter my own responses as i [I] have greater freedom and it's more testing of me".

In some of the cases, free-response questions were preferred because the students had issues with multiple-choice questions. Some students felt as if the multiple-choice questions were too easy, as the answer options provided could potentially lead the test-taker to the right answer:

"I preferred the questions in which you type some words. Multiple choice questions can often lead you to the right answer easily, but having to write out a response with no help makes you really question your understanding".

In contrast, other students felt that the options given by the multiple-choice questions were misleading, which could lead the test-taker to giving a wrong answer:

"The questions where I could type and answer as the multiple choice answers make me second guess myself".

Again, these reflections were similar to those given by participants who raised issues with multiple-choice questions through codes C3 and C4 in the usability laboratory study.

Findings from the Q34 responses: (b) Cases where students preferred multiple-choice questions

Out of the $N = 229$ students who gave responses to Q34 on Version 1 of the AMS, it was found that $N = 150$ preferred the multiple-choice questions. The different reasons for this preference are detailed in Table 5.3 below.

Reason for preferring MCQs	Tally	Meaning
Answer unambiguous	14	Refers to the student being able to select from a definite list of responses.
Easier/Quicker to answer	41	Refers to the multiple-choice questions being easier and quicker than their free-response counterparts.
Made them think	4	Refers to the students finding multiple-choice questions to be challenging to answer.
Provides guidance	38	Refers to the scaffolding provided in the multiple-choice questions to facilitate with giving an answer.
Writing own answers is hard	14	Refers to the student having difficulties when asked to write their own answer to the free-response questions.
Issues with the free-response question interface	11	Refers to the students having difficulties using the interface when answering the free-response questions.
No reason given	28	Refers to instances where the student stated that they preferred multiple-choice questions, but did not give any further explanation as to why.

Table 5.3: Reasons for students who did Version 1 of the AMS preferring the multiple-choice question type.

Some students liked the multiple-choice questions because they were quicker or easier to answer than the free-response questions. In contrast, other students found that the multiple-choice questions made them think carefully about the answers that they were giving:

"Select for [from] a list - Think about other answers in more depth and why they are right or wrong".

In other cases, students liked the idea of putting down an answer that was unambiguous:

"I preferred answering questions from a list of responses, as they provide an accurate interpretation of the intended answer".

Some students felt as if the multiple-choice questions provided them with guidance when they were stuck, and this was why they preferred them:

"Choosing from a list. I wasn't confident with all of my answers, so having options helped me to decide which answer was the most accurate as opposed to creating an

answer for myself, which I may have struggled to formulate".

Along the same lines, some students found that it was hard to construct their own answers to the free-response questions:

"I preferred the multiple choice because sometimes I didn't know what to write/ how to explain".

In other cases, the preference was pragmatic, as the students had encountered issues with the free-response interface which they did not with the multiple-choice questions.

Findings from the Q34 responses: (c) Cases where students preferred neither question type

Out of the $N = 229$ students who gave responses to Q34 on Version 1 of the AMS, it was found that $N = 35$ preferred neither question type. The different reasons for this preference are detailed in Table 5.4 below.

Reason for preferring neither question type	Tally	Meaning
Found both useful	19	Refers to the student identifying that multiple-choice questions and free-response questions can both be useful question types.
No reason given	16	Refers to instances where the student stated that they preferred neither question type, but did not give any further explanation as to why.

Table 5.4: Reasons for students who did Version 1 of the AMS preferring neither question type.

Some students found positives in both question types:

"I did not prefer either the multiple choice were made it easier to know you are giving the right answer whereas the ones that required words made you consider the problem in more detail".

In the other cases, students did not give a reason for not having a preference for either question type.

Combined discussion of AMS usability testing and Q34 response findings

From codes C1 and C2, the usability laboratory participants found that free-response format made them think about the questions, which allowed them to be creative when constructing their own answers. In this way, free-response questions were identified as being useful for students by making them think more deeply. These findings were backed up by the responses to Q34 given by students in the large-scale administration of Version 1 of the AMS, where free-response questions were identified as being useful for finding out about student understanding. Taken together, the above reflections showed that the students were capable of seeing the educational value of the free-response questions on the AMS. This finding illustrates that there is potential scope to include free-response questions in other concept inventories beyond the AMS.

From codes C3 and C4, participants identified that multiple-choice distractor options can lead test-takers to select an incorrect answer, making multiple-choice questions misleading; similar observations have previously been reported by Woodford and Bancroft (2005). In addition, some of the participants admitted to making use of *eliminate and guess* methods (Sangwin, 2013) and other strategic techniques when answering multiple-choice questions. Similar concerns were also raised in the responses to Q34, where multiple-choice questions were found to be confusing by some students; whereas other students thought that multiple-choice questions were too easy because of the guidance that is inherently built-in to them. Crisp (2007) pointed out that these factors make it difficult to draw conclusions about student understanding from multiple-choice questions, and this is one of the main motivations for making use of free-response questions in the AMS.

However, some participants noted that guidance was available in the question wording and answer options of multiple-choice questions, which could help them to answer the question when they were unsure. This contrasts with the above findings, where the same guidance was perceived as a negative aspect of the question type; this divide was also present in the Q34 responses. It is of note that students who were less sure of their answers were typically in favour of the guidance offered by multiple-choice questions, whereas students who wanted to challenge themselves by doing the questions tended to be against the guidance offered by multiple-choice questions. This indicates that the students had different perceptions of the multiple-choice questions based on what they wanted to get out of working through the AMS, and this could also be related to

the level of attainment of the students. Further data would be required to investigate this observation in a more general setting.

Somewhat surprisingly, evidence provided in Table 5.2 suggests that some students took longer to answer the multiple-choice questions than the free-response questions. It is possible that these students read through all of the possible multiple-choice options, which led them to do more on-screen reading in order to answer these questions; such complaints about computer-based assessment have previously been raised in the work of Nardi and Ranieri (2019). However, it is also possible that these students were reading through each of the options and evaluating the degree to which each option was wrong, rather than looking for the only option that could feasibly be correct. This is the reverse of the *eliminate and guess* strategies outlined previously, and instead illustrates how students could make use of the questions for learning purposes. In addition, students taking this approach to the questions were reading through them carefully, and this is a scenario which may be advantageous for female students (McBride, 2009; Dawkins et al., 2017).

In the case of code C5, usability lab participants and responses to the Q34 questions identified that multiple-choice questions were easier to answer, whereas free-response questions make students consider the problems in more detail. Thus, multiple-choice questions were perceived as being good for students, since they could answer these to build confidence; on the other hand, the free-response questions were perceived as being useful for the markers, because the answers better reflected students' understandings and misunderstandings. These findings highlight a contrast between what students and educators may perceive as positive aspects of question design. The multiple-choice questions can be perceived as positive by students because it makes their task easier; such a perception may not be shared by educators. Similarly, free-response questions can be perceived as positive by educators because they provide insight about students' conceptual understanding; this perception might not be shared by students. In this way, both free-response and multiple-choice questions have their merits for students and educators, and these need to be considered when designing questions of either type.

5.3.2 Findings related to the *Interpretations of the AMS instructions affected answer length* theme

There were 2 codes related to the *Interpretations of the AMS instructions affected answer length* theme, and these are presented in Table 5.5 below. This theme relates to the lengths of the answers to the AMS questions given by the participants, and it was coded 25 times overall. In addition, to facilitate the discussion of this theme, the mean answer length in words for each of the participants in the study is shown in Figure 5.4 below.

Code	Number of times coded
When responding to the information about answer length, some participants made their answers shorter (C6)	11
When responding to the information about answer length, some participants made their answers longer (C7)	14

Table 5.5: Codes associated with the *Interpretations of the AMS instructions affected answer length* theme.

The mean length of the answers to the free-response questions varied considerably by participant, as captured by codes C6 and C7. Participants P1, P5, P7 and P8 gave shorter answers, whereas participants P2, P3, P4 and P6 gave longer answers. Some of the reasons for this variation were explained by the participants' interview responses. Participants P1 and P3 both deliberated over the level of detail required in their answers, with P1 settling on giving shorter answers and P3 choosing to give longer, more descriptive answers:

"I felt that I'd better not just give a one-word answer, 'neither', although, I suppose that logically, that should have been adequate. But, I thought I'd better smooth it out a little bit" [P3] (P3 was referring to their answer to Q1. They did indeed type the full answer *"Neither - ignoring air drag, both should hit ground at same time"*).

Participant P2 claimed that they wanted to make their answers shorter; in spite of this, they typically gave longer answers in the form of sentences. Similarly, Participant P4 said that they attempted to modify the length of their answers as they worked through the AMS, but actually gave sentence answers in most cases.

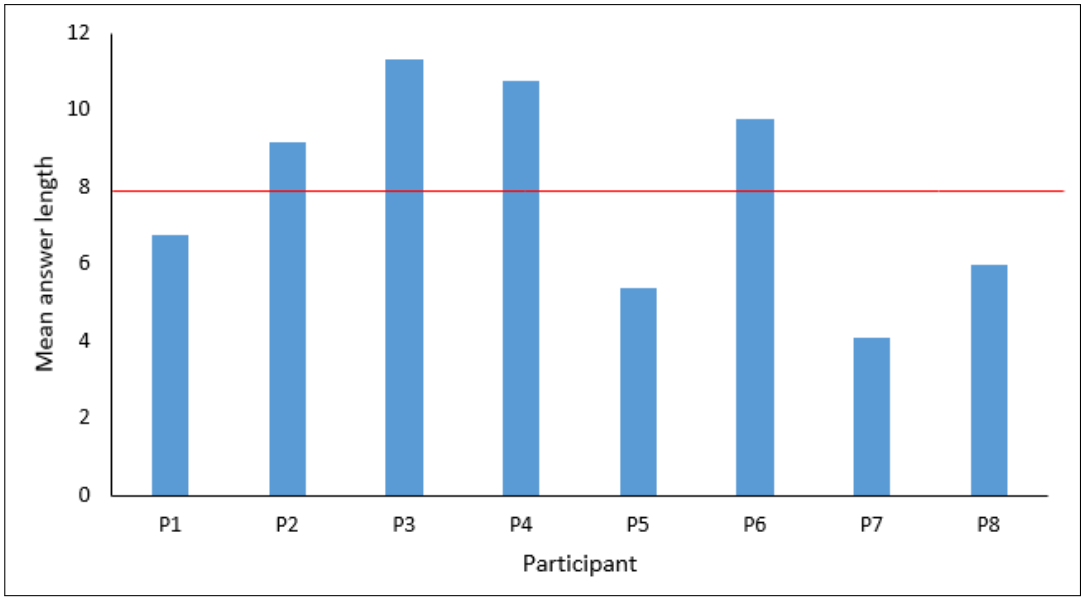


Figure 5.4: Graph showing the mean answer length in words by participant. The red horizontal line represents the mean of these mean answer lengths.

No participant attempted to submit an answer longer than twenty words. However, participants P5 and P7 used the information given in the instructions that answers of more than 20 words would not be accepted to guide their answer length. In the case of P5, the guidance made them feel as if the answers that they had given to the free-response questions were not detailed enough:

"I think I was worried about being either kind of vague, or, especially because there was a twenty-word limit, sometimes I felt that I would like to say a bit more, and was worried about how much depth I needed to go into with my response, I guess" [P5]

(P5 gave both short and long answers to the AMS questions. For example, they gave the one-word answer *"equal"* to Q20, but gave the longer answer *"when the top block moves between 3 and 4 at some point they will be the same speed"* to Q22).

On the other hand, P7 interpreted the guidance as advising them to write shorter answers to the free-response questions. In line with this, P7 gave mostly short answers, with several being only one word in length.

In summary, the length of participants' answers to the free-response questions were found to vary between participants. The data and responses suggested that none of the participants struggled with the length of their answers, but some were consciously aware of it. In addition, there was evidence that the participants used the guidance provided to determine what amount of detail was appropriate to present in

their answers. In the cases highlighted by code C6, the participants made their answers shorter, as the guidance had told them that answers of a few words would be sufficient; whilst in the cases identified by code C7, the participants made their answers longer, alluding to the 20 word limit that they had been alerted to. It is possible that the participants' reactions to the guidance were examples of regular exam techniques and study skills. However, it is also possible that these reactions were enhanced because the participants were entering their answers on a computer, rather than on paper. Further data pertaining to the length of hand-written student responses to a paper-based version of the AMS would be required to investigate this possibility.

Discussion of answer length by question

A consideration of the actual lengths of the answers given to the AMS questions is also relevant to the *Interpretations of the AMS instructions affected answer length* theme. The mean answer length in words for each free-response question on the AMS given by the participants in the study is shown in Figure 5.5 below. Figure 5.5 shows that Q17, Q22, Q30 and Q32 had the longest answers on average, whereas Q3 and Q25 had the shortest answers on average. The remaining 14 questions had short answer lengths, which is in keeping with the objective of having free-response questions that can be answered with a few words or a short sentence. As a further reflection of this, the overall mean answer length was 8 words, which is just under half the number of words permitted for each response.

The question with the shortest average length was Q25, which is worded as follows:

"As the rocket moves from position b to position c does the speed increase, decrease, or stay the same?"

This question is the second in a sequence of four questions based on the scenario of a rocket in space. Suggested answers are given within the question itself, and these suggested answers were used in all of the participants' answers. In contrast, Q30 was the question that had the longest answer on average. It is worded as follows:

"Now, the external force applied by the woman is suddenly switched off. Describe as accurately as you can what you think happens to the speed of the box after this event."

This question is the third in a sequence of three questions based on the scenario of a box being pushed. It instructed participants to describe a situation *"as accurately as they could"*, which is a likely explanation for the lengthy answers received.

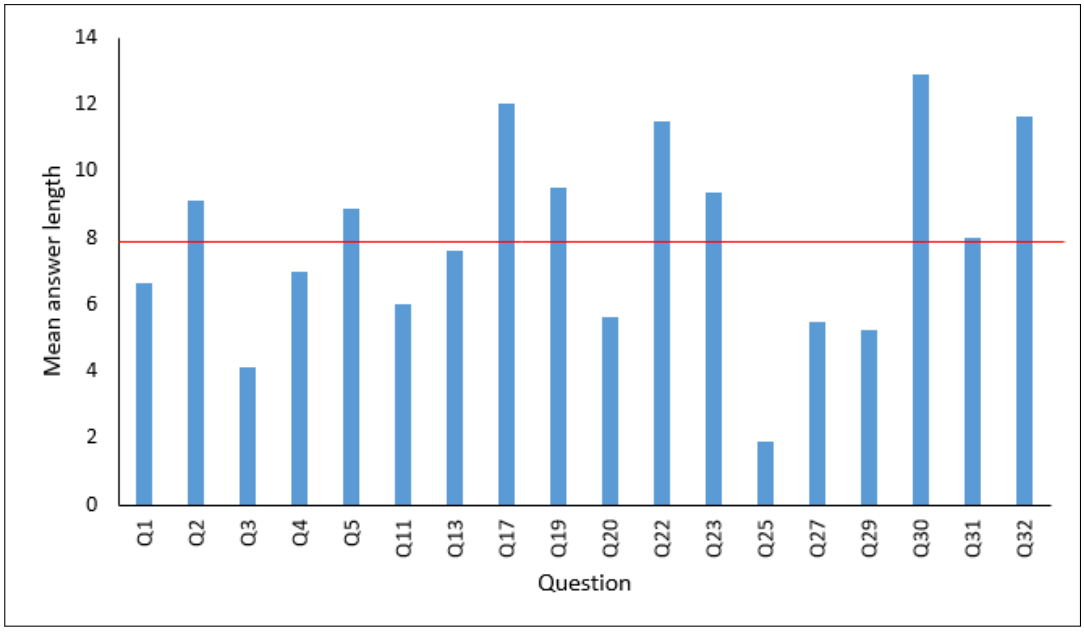


Figure 5.5: Graph showing the mean answer length in words by question. The red horizontal line represents the mean of these mean answer lengths.

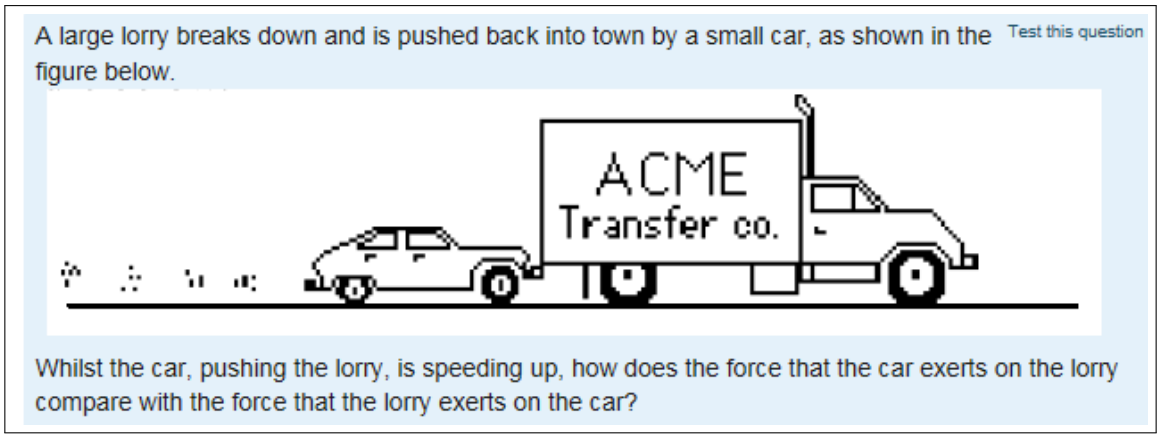


Figure 5.6: Q17 of AMS

Another question with longer answers was Q17, shown above in Figure 5.6. This question is the first of a pair of questions based on the scenario of a car pushing a truck. It is adapted from Q15 of the original FCI, which is considered in the literature to be a challenging FCI question (Scott et al., 2012), which may explain the length of the answers.

The above examples show that the participants gave answers based on the level of detail that they perceived the question to be asking them for. This is a demonstration of the participants acting as *conscientious consumers* (Higgins et al., 2002), doing

exactly what they think that they are supposed to do, even if that is not what was actually asked or intended. More generally, the impact of question wording on the responses given has implications for the design priorities for the authoring of questions in the future.

5.3.3 Findings related to the *The idea of being marked by a computer did not affect answer structure* theme

There was only one code associated with the *The idea of being marked by a computer did not affect answer structure* theme, and this code in given below in Table 5.6. The theme was coded 25 times overall, and it refers to the idea that the participants did not change the style of their answers based on the fact that they were being marked by a computer. Of note, P7 referred to this theme 9 times, which was more than the other participants.

Code	Number of times coded
When answering the free-response questions, some participants speculated about how the free-response questions were marked, but did not use this approach to try and beat the system (C8)	25

Table 5.6: Codes associated with the *The idea of being marked by a computer did not affect answer structure* theme.

From code C8, some of the participants thought about how the marking system worked for the free-response questions on the AMS, but none of them tried to use this sort of consideration in an attempt to *beat the system* while working through the AMS. In relation to this, P1 postulated that the computer was searching for keywords to mark the answers against, whereas P7 noted that it would be easier to setup effective automated marking for one-word answers than for sentence answers. In addition, no responses given by students to Q34 referred to using such approaches to answer the free-response AMS questions. This partially alleviates concerns about students trying to employ such approaches in a large-scale administration of the AMS, and providing skewed data. However, it is worth noting that none of the usability laboratory participants or student respondents to Q34 had any reason to be particularly invested in achieving a high score on the AMS, since it did not count towards any course grade. Further data would be required to investigate the prevalence of this

approach to answering free-response questions more generally.

The reaction of the participants towards the computer marking of their answers was indicative of the *Computers as Social Actors* framework (Reeves and Nass, 1996), which postulates that people can respond to a computer in the same way that they would respond to another person. It is of course important to ensure that answer matching of any free-response question is sufficiently accurate. Even in formative use, test takers are given a score which accurately reflects their level of understanding and accurate marking is important in order to retain student confidence. This was shown to be possible for short-answer free-response questions written with similar software (Butcher and Jordan, 2010), provided that student responses are used in developing the answer matching. In addition, particular care needs to be taken to ensure that alternative and incorrect spelling is accounted for, so as not to disadvantage users with dyslexia or for whom English is not their first language. These considerations highlight that free-response questions need to be carefully designed and tested in order to be deployed both inclusively and effectively in an educational setting (Jordan, 2012a; James, 2017).

5.3.4 Findings related to the *Participant reaction to the usability of the AMS was mostly positive* theme

The *Participant reaction to the usability of the AMS was mostly positive* theme was coded 57 times overall, and it pertains to the observation that the participants managed to work through the AMS without encountering issues most of the time. There were 3 codes associated with this theme, and these are outlined in Table 5.7. below.

Code	Number of times coded
In general, the participants found the AMS to be easy to use (C9)	27
There were cases where the participants encountered issues with the AMS interface (C10)	12
Participants were mostly against the idea of adding a time limit to the AMS (C11)	18

Table 5.7: Codes associated with the *Participant reaction to the usability of the AMS was mostly positive* theme.

Referring to code C9, all of the participants reacted similarly to the AMS, regardless of what their overall performance on it was. All of the participants read through the instructions at the beginning of the AMS, and all of them answered all of the questions. When faced with a free-response question for the first time, participants P1, P3, P7 and P8 read through the question first, before clicking into the text-entry box. On the other hand, participants P2, P4, P5 and P6 immediately clicked into the text-entry box when they saw it, and then read the question. These points demonstrate that in general, participants were able to use the interface and work through the AMS without difficulty.

In the cases of code C10, participants occasionally had problems with some of the finer details of the AMS. For example, participant P1 did not realize that they needed to scroll down the page containing Q3 to reach Q4; participant P2 felt diagrams would be useful in some of the questions that did not have them; and participant P3 had issues with answering Q3 because the dictionary did not recognize one of the words that they used in their answer. The AMS had a technical failure at one point for participant P6, and one of the observers had to enter the testing room and reset it manually.

Without a time limit in place, the participants worked through the AMS at their own pace, and took different amounts of time to answer the questions. P1, P2 and P5 took the longest time, with P2 taking 1 hour 16 minutes to complete the AMS, P1 taking 59 minutes, and P5 taking 52 minutes. P3, P4, P6 and P8 took less time, with P3 taking 39 minutes to complete the AMS, P6 and P8 each taking 36 minutes and P4 taking 33 minutes. P7 was particularly fast, completing the AMS in 22 minutes. Thus most participants took between half an hour and an hour to complete the AMS, and this information is given graphically in Figure 5.7 below. When asked to suggest a possible time limit for the AMS, participants P1, P4, P6, and P7 suggested various amounts of time, which are shown in Table 5.8 below.

Participant P1 suggested their time limit based on how long they thought the AMS would take to do on average. On the other hand, participants P4 and P6 estimated that since there were roughly 30 questions on the AMS, and if it took approximately 2 minutes to answer each question, then 1 hour would provide enough time to complete the AMS. Participant P7 instead suggested their time limit based on how long they took to complete the AMS, explaining why it is shorter than the other suggested times.

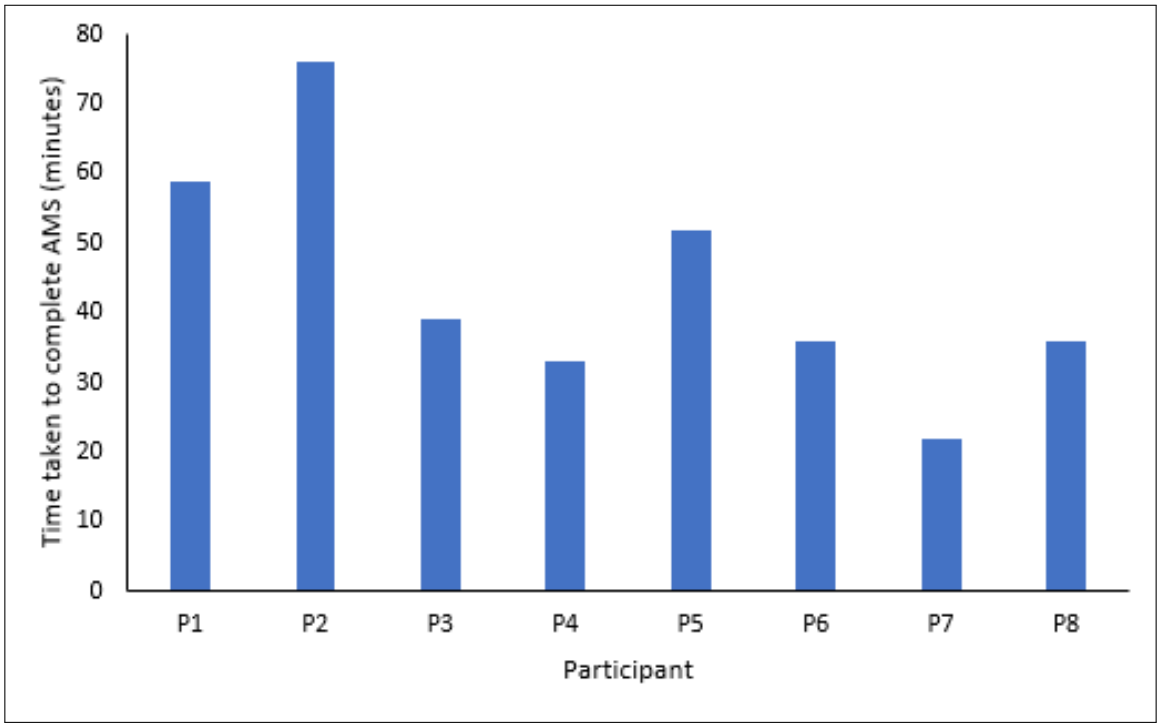


Figure 5.7: Graph showing the time taken to complete the AMS by each of the participants.

Participant	Suggested time limit
P1	45 minutes
P4	1 hour
P6	1 hour
P7	30 minutes

Table 5.8: Table showing the possible time limits for the AMS suggested by the usability laboratory study participants.

Since there was no time limit in place, no participant seemed rushed, and no other possible effects of adding time pressure emerged in the study. Instead, the participants highlighted some of the possible problems that the introduction of a time limit could bring, and these were gathered together in code C11. Participant P2 noted that adding a time limit would encourage students to answer the questions quickly; however, they also identified the drawbacks of this approach, since it would stop students from thinking carefully about the questions. Participant P6 felt that adding a time limit would put extra pressure on the test-taker, which could cause them to rush the questions and answer incorrectly. Participant P5 identified the key point that it is important to consider what the AMS was supposed to achieve before deciding whether to put a time limit on it:

"I think that depends on what you're planning to get out of it - Because it would then give somebody pressure, which would in some ways, stop you from thinking about what you are answering" [P5].

Implications for the development of the AMS

Occurrences of code C9 highlighted that participants were able to work through the AMS interface without issues most of the time. This was also found in the large-scale administration of Version 1 of the AMS, as students managed to enter responses to the questions, and there were no complaints about the usability of the AMS in the responses to Q34. When issues did arise, these were picked up by code C10; the issues highlighted by this code were mostly minor and question specific, meaning that they were easily resolved. Taking the findings from these codes together implied that the AMS was presented in a format that could easily be worked through by test-takers. This was a positive outcome from the usability testing, as it meant that the overall AMS did not need to be fundamentally re-designed for later versions.

Code C11 captured participants' ideas about adding a time limit to the AMS. Participants were not placed under any time limit when they worked through the AMS, and they each took different amounts of time to complete it. When asked, some participants suggested possible time limits for the AMS, and these were chosen from their own experience, or based on an estimate of the mean length of time taken by all participants. In general, the participants were not strongly in favour of adding a time limit, since it was not perceived to add anything useful to the AMS. However, some participants were against the idea of adding a time limit. These participants reasoned that if the AMS was designed to learn about student understanding, then adding a time limit would add unnecessary pressure to test-takers, which could lead to their levels of understanding being misrepresented. These reflections highlight the importance of considering what the purpose of the AMS is before trying to develop it further, since making unnecessary additions to the AMS could cause it to diverge from its original goals. This point is considered further in **Subsection 5.3.5**.

5.3.5 Findings related to the *Reactions to the AMS depended upon what participants thought it was for* theme

There were 2 codes under the *Reactions to the AMS depended upon what participants thought it was for* theme, and these are given below in Table 5.9. This theme was coded 44 times overall, and it relates to the ideas that participants had about what the purpose of the AMS was, and what it could potentially be used to do. It is of note that participants who were familiar with OU undergraduate study (so participants P1, P2, P3, P4 and P7) referred to this theme more often than the participants who were not familiar with OU undergraduate study (so participants P5, P6 and P8).

Code	Number of times coded
Within the assessment context, participants compared the AMS to the types of assessments with which they were familiar (C12)	19
Within an educational context, participants identified that the AMS could be used as a conceptual evaluation tool (C13)	25

Table 5.9: Codes associated with the *Reactions to the AMS depended upon what participants thought it was for* theme.

The frequency of codes C12 and C13 illustrated that the participants reflected upon the pedagogical purpose of the AMS. Participant P1 felt as if the AMS could be used to check that student understanding had reached a certain level, but could not be used to rank students based on their attainment. Participant P4 identified that it is important to consider whether the AMS would be used in a high-stakes or low-stakes situation when determining what its purpose would be:

"It depends on what the stake is; if this is, say, contributing to my overall score, or my moving forwards" [P4].

Further, the same participant noted the potential for the AMS to be used as a way of testing whether different teaching methods were effective or not, which aligns with the traditional use of concept inventories:

"It depends on what you're trying to do with the test - If the purpose of the test is, primarily for the teacher to work out whether or not the methods being used to teach the students are working" [P4].

Participant P7 considered how the AMS could be used within a course such as *S217*, feeling that the AMS would need to be made more advanced if it was to be used as an assessment at the end of a module:

"Well, it depends on what you're trying to achieve from it, really - If it is for an end-of-module exam, for whichever modules these things come in, then you probably want more actual numbers involved, and get them to do a bit of calculus" [P7].

The same participant also wondered what level of study the AMS would be appropriate for, believing that it would be suitable for first-year students. Note that in what follows, *level one* is the OU's equivalent of first-year university level study, and *level two* is the OU's equivalent of second-year university level study; in terms of UK for Framework for Higher Education Qualifications (UKFHEQ) levels, OU level one corresponds to UKFHEQ level four and OU level two corresponds to UKFHEQ level five:

"Is it for a level one module? - Perfect for a level one. For a level two? Probably not enough. Depends on what level you're aiming at" [P7].

Implications for future use of the AMS

Participants had various ideas about how the AMS could be used to assess students, and these were collected together by codes C12 and C13. Whether the AMS would be used in a high-stakes summative context or a low-stakes formative context was a key point raised by these codes. On the formative end of this spectrum (covered by code C13), one idea was that the AMS could be used to check that understanding of Newtonian mechanics had reached a certain level, which would make the AMS into a diagnostics test with similar objectives to the *Mechanics Diagnostics Test (MDT)* (Halloun and Hestenes, 1985) upon which the FCI was originally based. The summative end of this spectrum was covered by code C12. It was postulated that the AMS could be developed to test more advanced topics, or be used as part of an end-of module assessment. Beyond this spectrum, a further idea was that the AMS could be used as a way of testing the effectiveness of different teaching methods, which aligns with the traditional use of concept inventories (Porter et al., 2014). These reflections indicate that the AMS is perceived to be a versatile tool, with the potential for it to be used in both assessment and teaching contexts.

5.3.6 Findings related to the *Limited feedback was a useful addition to the AMS* theme

The *Limited feedback was a useful addition to the AMS* theme was coded 183 times overall, making it the theme that was coded most frequently. The theme refers to the limited feedback to the AMS questions that was given to the participants, and their reactions to this feedback. Note that the feedback is referred to as *limited* here since it only told the participants whether their answers were marked as correct or incorrect, with the correct answer also given to the multiple-choice and multiple-response questions. The theme consists of 4 codes, and these are shown in Table 5.10 below. Note that codes C14 and C17 were referred to particularly frequently by the participants.

Code	Number of times coded
Participants used the limited feedback provided by the AMS to reflect upon their performance (C14)	97
Some participants felt as if a greater level of feedback should be provided to the AMS questions (C15)	31
Some participants felt as if a lower level of feedback should be provided to the AMS questions (C16)	7
Participants responded positively to receiving feedback on their AMS performance (C17)	48

Table 5.10: Codes associated with the *Limited feedback was a useful addition to the AMS* theme.

Participants could access the feedback by clicking the *submit all and finish* button on the summary screen at the end of the AMS. A feedback screen then told them whether they had got each question right or wrong; for the multiple-choice and multiple-response questions, the feedback also gave the correct answer. All of the participants who received the feedback looked through their results, and used this to find out how well they had done. This showed that the participants were using the feedback to learn about their performance, and this behaviour was captured in the code C14, with examples given below.

On the feedback screen, participant P1 scanned through their answers, and looked for any instances where they were incorrect. They examined their three incorrect answers, Q2, Q17 and Q22, all free-response questions. They took some time deliberating over their answer to Q22, which asks the test-taker to identify the time interval where

a pair of blocks have the same speed, if at all. Participant P2 had several incorrect answers, and they scrolled through the marked questions, and looked carefully at their incorrect answers. Where there was feedback available, they read through it; where there was no feedback available, they re-read their answer and thought about it for a little while.

The author was in the room with participant P3 when they clicked on the *submit all and finish* button, and they asked for verbal explanations about their incorrect answers. This participant was also highly critical of the failings of the spell-checking system being unable to recognize some of the words that they had wanted to use, which turned out to be because the spell-checker was not active for non-keywords included in the answer. A related point was previously raised in **Subsection 5.3.3** where the importance of having accurate marking in order to give accurate feedback was alluded to. Participant P4 clicked on the *submit all and finish* button, and the author then went through the participant's incorrect answers with them. They also wanted verbal explanations about their incorrect answers.

After clicking the *submit all and finish* button, participants P5 and P7 both looked at the ticks and crosses given in the answer pane, and discussed their incorrect answers during the semi-structured interview. Participant P7 was particularly accepting of their own misconceptions when these were pointed out to them. Participant P6, who reported later that they were used to not getting feedback on examinations, did not click the *submit all and finish* button at the end of the AMS, so they were the only participant to not receive the feedback.

Participant P8 looked immediately at the answer pane, and counted up how many questions they had got right and wrong. During the interview, they asked about many of the questions, and sought clarity on question wording and meaning, as well as what the correct answers were. Participant P8 used the experience to try and learn about gaps in their knowledge, and reported that verbal feedback was useful to them in this respect. They were often frustrated, apparently because they felt that they were capable of answering several of the questions correctly, but had not. It is of note that participant P8 was attentive to any advice given when they answered a question incorrectly.

The participants had different views on how much feedback should be given, which was picked up through codes C15 and C16. Participant P1 felt as if being told whether

they were right or wrong was sufficient, since they could tell what they had done wrong from this low-level feedback, and then take the necessary steps to improve without further prompting. However, participants P2 and P5 noticed that no model answer was given to the free-response questions, and they felt as if more detail should be provided in the feedback to these questions for guidance:

"I was expecting something for the typed questions...it would have been good to have the feedback" [P2].

On the other hand, participant P7 instead thought that giving model answers in the feedback might discourage students from going back and revising the material properly. Instead, they suggested that the feedback should guide the test-taker to the relevant section of the study material:

"I mean, you could, have a little bit that says 'refer to section whatever of which book', that might be a half-way house" [P7].

Participant P3 reasoned that the level of the topic being tested should be used to determine the level of detail of the feedback provided. For instance, this participant felt as if the level of feedback given by the AMS was appropriate, but more feedback would be required for higher-level topics:

"Well, it's fairly low-level, let's be clear on that. If it were a more advanced topic, I would probably have required, I would have welcomed, even more so, a more detailed feedback" [P3].

Related to this, participant P4 pointed out that the level of feedback required also depended upon the purpose of the AMS. As a result, participant P4 postulated that if the AMS was meant as a diagnostic test, then summary feedback with a list of topics requiring attention would be helpful:

"A test that tests similar concepts throughout in different ways - and at the end, I didn't have to have a think - and it just told me a score and told me that 'the areas that you were weak on were Newton's Third Law, second law', or whatever - then I also think that that would be sufficient" [P4].

The participants generally reacted positively to receiving feedback from the AMS, as highlighted by the occurrences of code C17. For participants P1, P2, and P5, this was because the feedback told them how well they had done:

"Well, I wouldn't have enjoyed not getting any feedback. I think it's the same with anything that you do; if you do an exam, you want to know how well you did" [P1].

For participant P4, the feedback was welcomed because it told them where they had gone wrong, which allowed them to improve their understanding by referring back to the relevant course materials:

"I think it's helpful, or it will help me now to kind of target my results - which were [pause] more disappointing" [P4].

Discussion of giving feedback to the AMS questions

When the participants did the AMS, the instructions did not tell them that they were going to get feedback on their performance, or what the detail of this feedback would be. All but one of the participants received the feedback, and the behaviour and interview responses of these participants indicated that they were not surprised to get some sort of feedback after completing the AMS. Participants who received the feedback were found to be interested in their own performance on the AMS, as captured by code C14. All the participants who received the feedback were observed to scroll through their answers to check where they were right and where they had gone wrong; in general, participants paid more attention to the instances where they had been wrong. In these cases, participants analyzed their own answers and asked questions to the interviewers about why their line of reasoning was incorrect. As a result, these participants were interested in using the AMS usability testing experience to better their understanding of the concepts being assessed. This sort of self-regulated learning (Nicol, 2007) gives students responsibility for their own learning. It is a well-established principle in formative assessment, but worthy of further investigation in the context of concept inventories; this is because concept inventories do not usually give feedback to students, although this has been done occasionally with the aim of increasing student self-efficacy (Chen et al, 2004; Lawrie et al., 2013).

Participants had different ideas about the level of the feedback that should be given by the AMS, and these were encapsulated within codes C15 and C16. At this point, it is important to note that levels of feedback on assessed tasks can be classified in several different ways. Carless (2006) considers whether the purpose of the feedback is advice for improving the current assessment, advice for future assessments, a means of explaining or justifying a grade, or a ritual. Shute (2008) instead classifies the different

types of feedback that can be given as verification of response accuracy, explanation of the correct answer, hints, and worked examples.

In line with Weaver (2006), some of the participants felt as if being told whether they were right or wrong was sufficient, because they could tell what they had done wrong from this low-level feedback, and then take necessary steps to improve without further prompting. However, in a concept inventory, this approach assumes that students can work out the nature of their conceptual misunderstanding without further guidance. To counter this possible issue, other participants felt as if a model answer should be given, which would serve to highlight where their line of reasoning had gone awry. However, a potential disadvantage of this approach is that students could memorize the answers for future use, rather than building up their physics knowledge and understanding. Such behaviour has been reported previously (Bull and McKenna, 2004), although where students were encouraged to take responsibility for their own learning, this behaviour was found to be generally limited to situations in which students found the question or the feedback difficult to understand (Jordan, 2009).

One participant reasoned that the level of the topic being tested could be used to determine the level of the feedback provided. This participant felt that the level of feedback given by the AMS was appropriate, but a higher level of feedback would be required for more challenging topics. Related to this, another participant pointed out that the level of feedback required was related to the purpose of the AMS. As an example, the participant postulated that if the AMS was meant as a diagnostic test, then summary feedback with a list of topics requiring attention would be helpful. These participants' ideas about different levels of feedback are comparable to Nyquist's distinction of *weaker feedback*, in which students are just told about their score, as compared with *stronger formative assessment*, which gives information about correct answers, explanation of the answers, and activities to undertake to improve (Nyquist, 2003).

Getting some feedback was perceived as an important part of the process by most participants, and this was highlighted by code C17. The feedback was found to be useful by the participants, even though it was limited to knowledge of whether their responses were right or wrong, with the correct answer also given for some of the questions. This is in agreement with the findings from the literature that students like receiving feedback, even if they do not properly make use of it (Brown and Glover,

2006) or if the feedback intervention is actually unhelpful (Kluger and DeNisi, 1996).

For some participants, feedback was seen as positive because it told them how well they had done. When making use of limited feedback in this way, students may simply be checking that they are making reasonable progress, in line with the findings of Scott (2014) and Draper (2009). At an even lower level of feedback, Millar (2005) found that students are interested in knowing their score, even if this does not contribute in any way to their course grade, as was the case for the AMS. For another participant, feedback was welcomed because it showed them where they had gone wrong, allowing them to refer back to the course material to improve their own understanding. As was the case for this participant, the presence of computer-generated feedback has previously been shown to deter students from using a trial and error approach to answering the questions (Walker et al., 2008), which ties in with the aims of using concept inventories to investigate student understanding.

Whatever their reasons for wanting feedback, and whatever use they saw that it had, the participants in general saw feedback as a good thing. Giving feedback to students enables them to take responsibility for their own learning and allows them to gain independence (Boud and Soler, 2016). Meanwhile, the more conventional use of concept inventories to provide feedback to teachers is in line with the recently recognized field of *learning analytics* (Sedrakyan et al., 2018; Zilvinskis et al., 2017; Clow, 2013). The provision of feedback to both students and teachers marks a welcome move towards a more personalized type of teaching and learning, where students' needs are responded to in a way that is based upon their own strengths, weaknesses and willingness to engage. In the context of the current work, this could be an area for further investigation in the future.

5.4 Conclusions

The first aim of the study was to investigate how students reacted to free-response concept inventory questions; the second aim of the study was to investigate how students reacted to being given feedback on concept inventory questions. Data were collected for the study by having eight participants work through the AMS in a usability testing setting, and conducting interviews with the participants about their experience. Further data were collected from the large-scale administration of Version 1 of the AMS in the form of qualitative responses to the feedback question Q34, and the findings

from the different data sets were triangulated where relevant.

In the context of the first aim of the study, participants generally reacted positively to being asked AMS questions in the free-response format. It was found that free-response questions made participants think more deeply about the questions, which encouraged them to be creative when writing their answers. Participants also noted that answers to free-response questions provide more information about student understanding to the marker; in this way, participants could see the educational value of using free-response questions instead of multiple-choice. Taken together, this suggests that it is feasible to use free-response AMS questions in place of the multiple-choice FCI counterparts, which validates their use in the AMS.

In the context of the second aim of the study, participants viewed getting feedback as an important part of the process of working through the AMS, and they responded well to receiving it. Feedback was found to be useful by participants because they were interested in finding out about how they had done on the AMS. The feedback was limited in detail, and participants had different ideas about the level of feedback that should be given by the AMS; these were often related to what the participants thought that the purpose of the AMS could be. Taken together, this suggests that there is an opportunity to make use of formative concept inventories that give feedback as a tool for guiding more independent, student-driven learning.

5.5 Summary and looking ahead

Chapter 5 presented the qualitative findings from the usability laboratory study and the responses to Q34. It was found that participants could see the educational value of using free-response questions instead of multiple-choice, and they responded well to these questions. In addition, students welcomed feedback after working through the AMS, and most were seen to make use of it.

Chapter 6 focuses on testing the AMS questions and marking rules using quantitative approaches. It presents findings from data gathered through administration of Version 1 of the AMS in the academic year 2017-2018.

6 Applying free-response questions to the Alternative Mechanics Survey

6.1 Rationale

Chapter 5 examined how students reacted to the free-response *Alternative Mechanics Survey* (AMS) in a usability laboratory study. It was found that students generally responded to the questions in the expected way, meaning that they were testing what they were expected to test, which gave a qualitative validation for the use of free-response AMS questions. However, the questions also needed to be tested for *reliability* (meaning that they produce consistent results), and this required a quantitative approach. Within the literature, a commonly used approach to such testing is *Classical Test Theory* (CTT). In addition, the marking rules needed to be tested for effectiveness, and this also required a quantitative approach. Within the literature, the *Inter-Rater Reliability* (IRR) approach is often used for this testing. In the work presented in this chapter, these two quantitative analysis approaches were applied to the data set collected using Version 1 of the AMS; the results were used to test the reliability of the AMS questions and marking rules as well as allowing a more detailed investigation into the automated marking of the free-response questions in the AMS.

6.2 Methods

6.2.1 Data collection

The Version 1 AMS data set was collected by putting the Version 1 AMS questions into a Moodle test hosted on the OpenScience Laboratory (OSL). After gaining the relevant approvals from The Open University's *Human Research Ethics Committee* and *Student Research Project Panel*, data could be collected. This was done by contacting potential participants with information about the project, together with a link to Version 1 of the AMS on the OSL. The potential participants contacted were: Open University undergraduate students on the modules *S112 Science: Concepts and Practice* (*S112* is a Level 1 interdisciplinary science module), *S383 The Relativistic Universe*, *SM358 The Quantum World*, and *SMT359 Electromagnetism* (*S383*, *SM358* and *SMT359* are Level 3 physics modules); undergraduate physics students from the University of Edinburgh; and secondary school students contacted via their teachers. The data were downloaded from the OSL once all of the participants had responded to the AMS. Blank entries were removed, and complete tests were retained for calculation of the

CTT statistics; all non-blank entries for each question were separately retained for calculation of the IRR statistics.

For the IRR calculations to be done, human marking of the responses was required to give the computer marking something to be compared against. However, humans do not mark in a consistent way, even when provided with a mark scheme (Butcher and Jordan, 2010), so more than one human marker was required in order for the comparison of human and computer marking to take place. For this task, five human markers were recruited from The Open University's School of Physical Sciences. All of the markers were staff members with a background in physics, so the subject matter of the AMS should have been straightforward for them to comprehend and mark.

For the responses to Version 1 of the AMS, the same set of marking guidelines was given to each of the markers. The markers were instructed to award a mark of 1 for answers that they deemed to be *correct* and a mark of 0 for answers that they deemed to be *incorrect*, with no partial credit given. The marked responses for each question from each marker were collected together to establish a *Unified Human Marker* **(UHM)** by examining how the majority of the markers chose to mark each of the responses. For example, if a response was marked as correct by 4 of the human markers, and incorrect by 1 of the human markers, then the UHM would award a mark of 1 for this response, being the majority view. Since it is built from a consensus of experts, the UHM is able to provide the most reliable marking for each response; hence the UHM was treated as the *master mark scheme* in the study.

Cases where the response was marked one way by 3 of the markers and the other way by 2 of the markers were deemed to be *borderline*, and were re-examined by the author and the lead supervisor. Marks in the borderline cases were only changed if they were inconsistent with other similar cases within the UHM. The final version of the UHM was then used to test the accuracy of both the computer's marking and the accuracy of the the individual human markers.

6.2.2 Data analysis

Classical Test Theory

Classical Test Theory (CTT) is outlined in detail in the pioneering work of Crocker and Algina (1986), and is described in less detail in works such as Ding and Beichner

(2009). CTT assumes that the total test score observed when a student does a test is the sum of a *true score* that corresponds to the student's actual level of understanding, and a *guessing score* that corresponds to the student correctly guessing the correct answer. This is represented in the following equation:

$$X = T + E \tag{6.1}$$

where X is the total observed score, T is the true score attained through understanding, and E is the error score invoked by guessing. With this equation as the starting assumption, CTT outlines statistics that can be used to test various aspects of the test's functionality. Some of these statistics are calculated for individual items, whereas others are calculated for the entire test. These statistics are described here.

The *difficulty* of a test item is the proportion of test-takers who, in completing all questions in the test, answered the test item correctly. It is calculated using the formula:

$$P = \frac{N_1}{N} \tag{6.2}$$

where P is difficulty, N_1 is the number of correct responses, and N is the number of test-takers. A larger difficulty value therefore corresponds to an easier question. The difficulty takes a value between 0 and 1, with the acceptable range of values for difficulty being from 0.3 to 0.9 (Doran, 1980). Difficulty values that are below the 0.3 cut-off are considered to be too hard, whereas difficulty values that are above the 0.9 cut-off are considered to be too easy; items with difficulty values outside of the range are candidates for revision. A difficulty of 0.5 represents a perfectly balanced item, although it is desirable to have a range of different difficulty values across a test to help it to differentiate between higher and lower scoring test-takers. Since difficulty is not dependent on the overall scores of the test-takers, or on the structure of the overall score data, it can be calculated individually for each question based on the number of respondents who gave an answer to the question; in this thesis, this is referred to as the **dynamic difficulty**.

The *discrimination* is defined as the item's ability to distinguish between higher-performing test-takers and lower-performing test-takers. It is calculated using the formula:

$$D = \frac{N_H - N_L}{N/4} \tag{6.3}$$

where D is discrimination, N_H is the number of correct responses from test-takers in the upper quartile of overall test score, N_L is the number of responses from test-takers in the lower quartile of overall test score, and N is the total number of test-takers. The discrimination takes a value between 0 and 1, with the acceptable range of values for discrimination being from 0.3 to 1 (Doran, 1980). A larger value of the discrimination corresponds to a question that is more effective at distinguishing between test-takers of different performance levels, thus a larger value for the discrimination is preferred. When an item has a discrimination value that is outside the acceptable range, there may be some issue with the clarity of the question wording that is causing it to be a poor discriminator.

A statistic related to the discrimination is the *point biserial coefficient*. The point biserial coefficient is a measure of the correlation between the scores on the item and the total scores for the entire test. It hence measures how well the item tests material that is consistent with the rest of the test. It is calculated using the formula:

$$r_{pbi} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_x}\sqrt{P(1-P)} \tag{6.4}$$

where $\bar{X}_1$ is the mean total test score of the test-takers who answered the item correctly, $\bar{X}_0$ is the mean total test score of the test-takers who answered the item incorrectly, σ_x is the standard deviation of all of the scores, and P is the difficulty of the item. The point biserial coefficient takes a value between 0 and 1, with the acceptable range of values being from 0.2 to 1 (Kline, 1986). A higher value of the point biserial coefficient corresponds to a question that tests material that is more consistent with the rest of the test, thus a larger value for the point biserial coefficient is preferred. On the other hand, when an item has a point biserial coefficient value that is outside the acceptable range, the item may not be testing content that is the same as the rest of the test, or is not on the same level as the rest of the test.

The entire test will be *reliable* when it is consistent. This means that if the same test-takers did the same test repeatedly without learning from the experience, they would be expected to get the same scores. This is not a feasible experiment to conduct, so reliability needs to be treated in a different way. If the items test similar material, test-takers would be expected to give similar responses on these items. With

this is mind, the *Kuder-Richardson reliability* measures the extent to which a test is constructed using questions that test similar materials. It is calculated using the following equation:

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum P_i(1-P_i)}{\sigma_x^2}\right) \tag{6.5}$$

where K is the number of test items, P_i is the difficulty of the i^{th} item, and σ_x is the standard deviation of the total score. A higher r_{test} value indicates that the questions test similar material, making the test overall more reliable. An r_{test} value of 0.7 is considered to show that a test is reliable overall, with higher values being better still.

The Kuder-Richardson reliability expands the idea of the point biserial coefficient testing reliability of individual test items to testing the reliability of an entire test. In a similar way, *Ferguson's delta* expands the idea of the discrimination coefficient to assessing the discrimination of an entire test. It is calculated using the following equation:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - N^2/(K+1)} \tag{6.6}$$

where N is the number of test-takers, f_i is the number of test-takers who scored i on the test, and K is the number of items on the test. A δ value of 0.9 is considered to be acceptable, and shows that the overall test has good discriminatory capabilities.

These five statistics form the basis of the CTT analysis used to test the AMS questions, and calculations of them for the Version 1 AMS data set can be found in **Section 6.3**. Before detailing these results, the Inter-Rater Reliability statistics used to test the AMS marking rules are introduced.

Inter-Rater Reliability

There are various statistics that can be used to calculate the Inter-Rater Reliability (IRR), and the appropriate statistic must be chosen based on the properties and assumptions of the situation. Some such statistics are outlined by Artstein and Poesio (2008), as well as by Zwick (1988). The most basic such statistic is known as the *percentage agreement*, which measures the proportion of cases where the raters agree on the classification of subjects. It is calculated as follows.

The agreement value agr_i for the subjects i is defined as:

$$\text{agr}_i = \begin{cases} 1 \text{ if the two raters assign } i \text{ to the same category} \\ 0 \text{ if the two raters assign } i \text{ to different categories} \end{cases} \tag{6.7}$$

Percentage agreement A_0 over all of the subjects i is hence:

$$A_0 = \frac{1}{n} \sum \text{agr}_i \tag{6.8}$$

where n is the total number of subjects classified. Percentage agreement takes a value between 0 and 1, with values of 0.95 and above indicating a good level of agreement (Jordan, 2012b). However, a high percentage agreement value alone is not sufficient to show that agreement is genuinely at a good level, because percentage agreement does not take into account sample size or chance agreement. As its name suggests, chance agreement is agreement that arises when raters assign a subject to the same category by random chance. More advanced IRR statistics are designed to account for random chance, as is outlined below.

A_0 was already defined as the value of the percentage agreement; A_e is now defined as the agreement that is expected to arise by chance. Therefore, the value of $1 - A_e$ gives the maximum amount of *true agreement* (not by chance) that is possible to attain, whereas the value of $A_0 - A_e$ gives the amount of *true agreement* that is observed. Dividing $A_0 - A_e$ by $1 - A_e$ therefore gives the proportion of *true agreement* that was observed, hence accounting for the agreement that arises by chance. The agreement statistic when chance agreement is accounted for is defined as:

$$A = \frac{A_0 - A_e}{1 - A_e} \tag{6.9}$$

The advanced IRR statistic that follows is calculated using this formula for A. In particular, the way in which A_e is calculated for this statistic based on the assumptions made about the way that raters classify subjects is explained.

Cohen's kappa (κ) (Cohen, 1960) assumes that the raters would get different distributions if they did their classification by chance. This is a realistic scenario with respect to the current study, since raters would not be expected to do their classifications in the same way, even when doing so at random. Mathematically, this means that if the raters are using categories labelled k, whilst $n_{c_1 k}$ is the number of times that the first rater assigns an object to category k, $n_{c_2 k}$ is the number of times that the

second rater assigns an object to category k, and j is the number of objects classified, then the probability P_{c_1k} of the first rater assigning an arbitrary object to category k is given by:

$$P_{c_1k} = \frac{n_{c_1k}}{j} \tag{6.10}$$

Similarly, the probability P_{c_2k} of the second rater assigning an arbitrary object to category k is given by:

$$P_{c_2k} = \frac{n_{c_2k}}{j} \tag{6.11}$$

Hence the probability P_k of both raters assigning an arbitrary object to category k is given by:

$$P_k = P_{c_1k} \times P_{c_2k} = \frac{n_{c_1k}}{j} \times \frac{n_{c_2k}}{j} = \frac{n_{c_1k}n_{c_2k}}{j^2} \tag{6.12}$$

Summing over the K different classifications gives the A_e value for Cohen's κ of:

$$A_e^\kappa = \sum \frac{n_{c_1k}n_{c_2k}}{j^2} = \frac{1}{j^2}\sum n_{c_1k}n_{c_2k} \tag{6.13}$$

Putting the A_e^κ from equation (6.13) as the A_e value in equation (6.9) gives the corresponding Cohen's kappa (κ) statistic, which takes a value between 0 and 1. Values for Cohen's kappa that are 0.8 and above are considered to be acceptable (Artstein and Poesio, 2008), and illustrate good rater agreement. Cohen's kappa is used in this study because its assumption that different markers will have different marking distributions even when marking at random, matches with the expected behaviour of human and computer marking. In addition, the percentage agreement is used in this study, although it is referred to as the *marking agreement* as a result of the automated marking context. As previously noted, the marking agreement takes a value between 0 and 1, with values of 0.95 and above considered to be acceptable. In addition, since the marking agreement does not take into account chance agreement, it provides an over-estimate of the agreement between the markers. However, the marking agreement is used in this study because it provides a rough idea of how well (or not) the computer marking is functioning on a given question, as well as giving a base value to compare the corresponding Cohen's kappa statistic against.

These IRR statistics were used in this study as follows. For each of the free-response questions, the marking agreement and Cohen's kappa values were calculated for the UHM against the computer marker; these values were then used to identify problematic cases where the computer marking rules were not functioning at the desired level, as well as to identify generic difficulties in the computer marking. The number of times the UHM disagreed with the computer marker on each question was also noted, and the number of cases which were *false positives* and *false negatives* within these cases were also noted. The cases on each question were used to improve the marking rules by adding suitable negation rules to counter the false positives; and by using the false negative cases to cover other correct answers. To check for consistency, these new marking rules were also back-tested against the responses used to develop them, as well as against other previous response sets where appropriate. This process for developing and testing the marking rules is illustrated in Figure 6.1 below.

A similar approach has previously been used successfully by Butcher and Jordan (2010) to develop and improve marking rules for a similar automated marking system. The approach does have known limitations; for example, Butcher and Jordan recognized that some false positive cases are difficult to resolve because the addition of negation rules can inadvertently give rise to new false negative cases. Furthermore, trying to add marking rules to account for every different false negative answer wording can lead to the occurrence of an over-fitting problem (Zehner et al., 2016). These issues are discussed in more detail within the context of the current study in **Subsection 6.4.2**.

In addition the performance of the UHM and individual human markers were tested by calculating the marking agreement and Cohen's kappa values for the UHM against each of the individual markers. This additional testing served two purposes: first, it tested the UHM for internal consistency; second, it showed the extent to which the human markers were inherently inaccurate, and provided a way of investigating why this inaccuracy arose.

In what follows, findings from the Version 1 CTT analysis are presented and discussed in **Section 6.3**, with findings from the Version 1 IRR analysis similarly detailed in **Section 6.4**. The significance of these findings for the development process of the AMS are highlighted afterwards, in **Section 6.5**. The Version 1 AMS questions used to conduct these studies can be found in **Appendix C**.

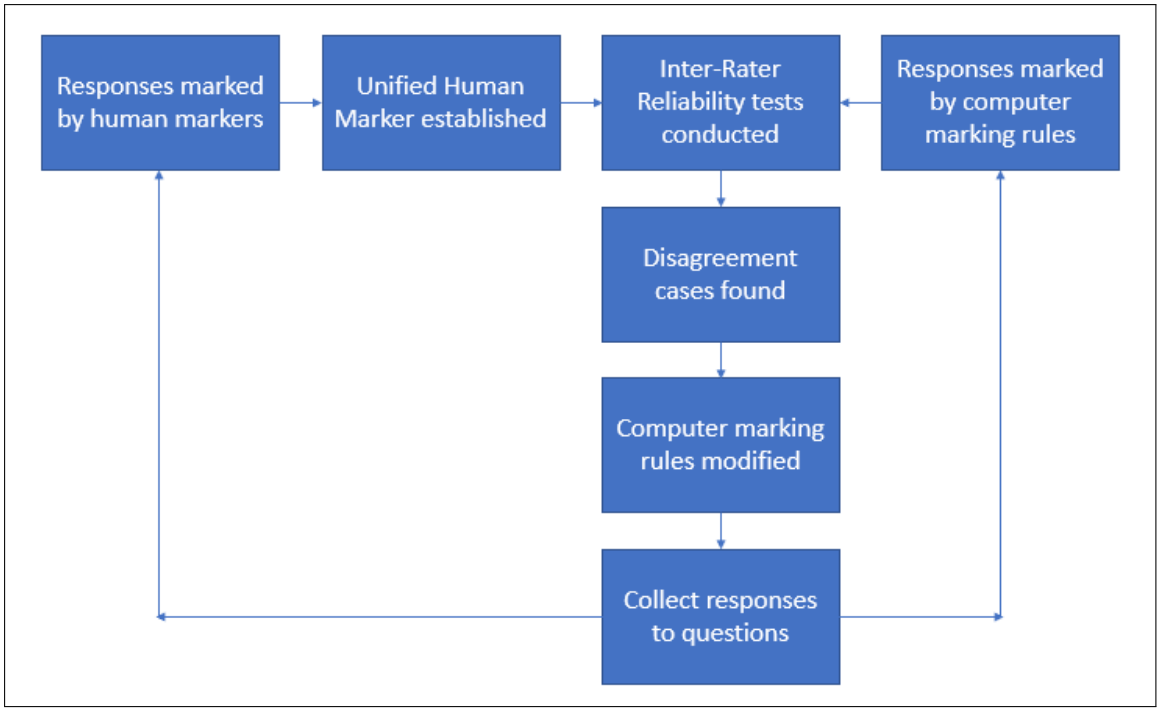


Figure 6.1: Flowchart illustrating the IRR-based process used to test and develop the AMS computer marking rules.

6.3 Results and Discussion: AMS Version 1 CTT study

6.3.1 Total score and number of attempts

There were 328 respondents to Version 1 of the AMS (Of these, 145 were high school students, 148 were undergraduate students, and 35 were STEM faculty staff at the OU), and 254 submitted tests which were *complete*, meaning that the respondents had answered all 33 of the questions (Of these, 122 were high school students, 105 were undergraduate students, and 27 were STEM faculty staff at the OU). The graph showing the frequency of each of the different scores for the $N = 254$ completed tests is given in Figure 6.2 below. Note that these are the scores as were awarded by the Version 1 UHM.

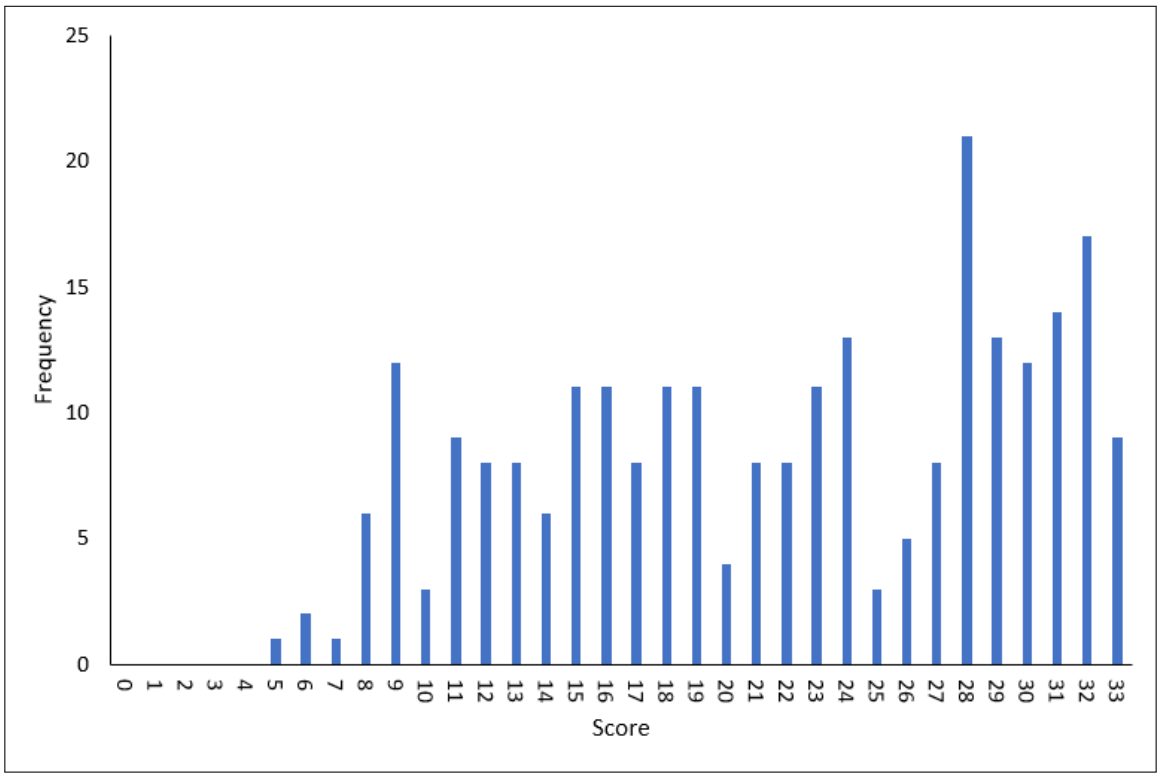


Figure 6.2: Graph showing the frequency distribution of the Version 1 AMS overall test scores for all 254 completed tests marked by the UHM.

Figure 6.2 shows that the modal score on the AMS was 28. The scores do not follow a **normal distribution**. However, the shape of the graph could possibly be built up from separate normal distributions associated with the various populations of test-takers. In addition, the mean score on the test was 21.60, which is above the middle value of 16.5. This is significant because if each item on the AMS had the optimal difficulty of 0.5, then the mean score would be expected to be 16.5, which is half-marks.

The data is left-skewed, and this can possibly be explained by the demographics of the participants who took the AMS. The respondents were made up of high school students, conventional undergraduate students and distance learning undergraduate students. The AMS was presented to these students as an optional activity, and no extra credit or incentives were offered for participation. This means that those who engaged with the AMS and completed it were likely to be only the most enthusiastic students (Hunt and Jordan, 2016), who are also often the same students who know the content well. Additionally, all of the participants will have already encountered Newtonian mechanics as part of their education, so the data gathered can be considered to be post-test. Given this information about the participants, a higher average score would be expected on the AMS.

Since test-takers had the option of abandoning the AMS at any time, 74 of them did not complete the AMS. Figure 6.3 shows how far through the test each the 328 respondents got before giving up on the AMS. Note that $N = 254$ respondents completed the entire AMS.

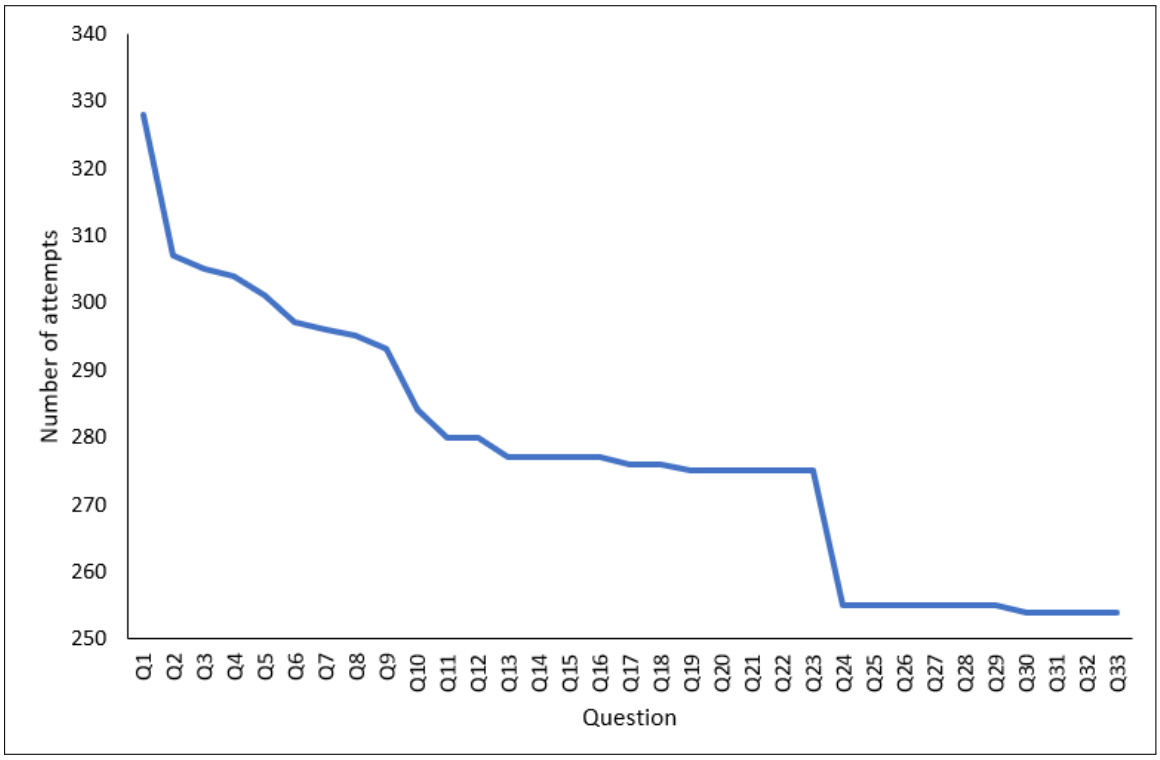


Figure 6.3: Graph showing the number of attempts made on each question on Version 1 of the AMS.

Figure 6.3 shows that there was a sharp decrease in participation from Q1 to Q2. This was possibly because respondents looked at Q1 and attempted it, before deciding to not proceed with the remainder of the test. There was then a gradual decrease until Q13, where the graph leveled out. After, there was one more sharp decrease between Q22 and Q23, after which it again leveled out. Unlike the situation for the drop in engagement between Q1 and Q2, no obvious explanation has been found for the drop after Q23, which indicates that the effect in this case may have had something to do with the properties of Q23. This question required test-takers to identify if either of a pair of moving blocks is accelerating, and it corresponds to Q20 of the original FCI. It is shown in Figure 6.4 below.

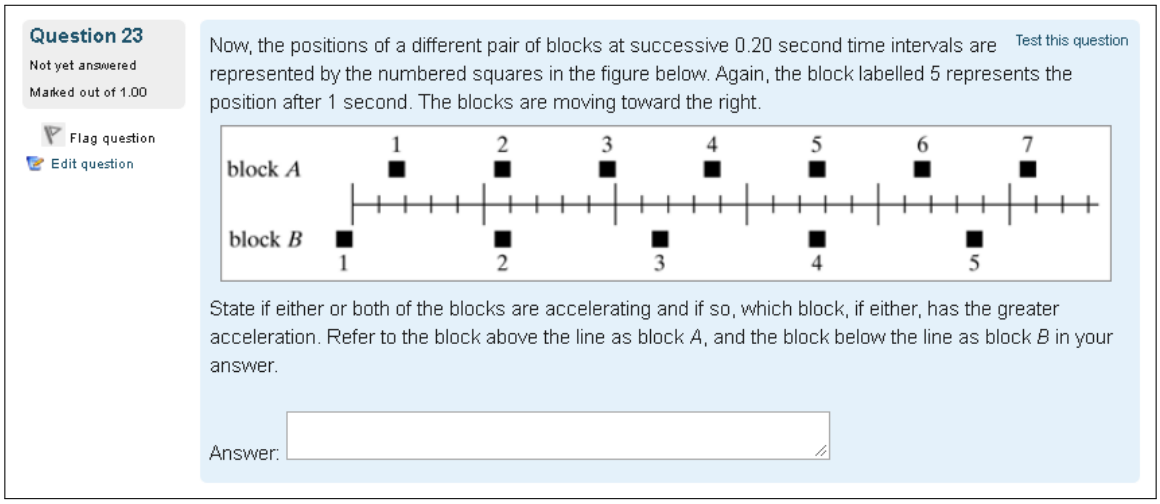


Figure 6.4: Q23 of Version 1 of the AMS, which is adapted from Q20 of the FCI.

Marin-Blas et al. (2010) found that students with a lower previous exposure to physics were less than half as likely to get FCI Q20 correct than those with a higher exposure to physics. Applying this finding to the drop-off observed on Q23, it is possible that students with a lower previous exposure to physics found this question particularly difficult, and decided to give up on the rest of Version 1 of the AMS as a result.

6.3.2 Difficulty and dynamic difficulty

Two versions of the difficulty statistic are calculated here: the *difficulty* and the *dynamic difficulty* as previously defined in **Subsection 6.2.2**. Data pertaining to these two types of difficulty are presented in Table 6.1 and Figure 6.5 below. Note that the data used for these calculations were the Version 1 AMS responses marked by the UHM. Further note that in Table 6.1 and subsequent tables in the thesis, FRQ denotes that the question was free-response; MRQ denotes that the question was multiple-response; and MCQ denotes that the question was multiple-choice.

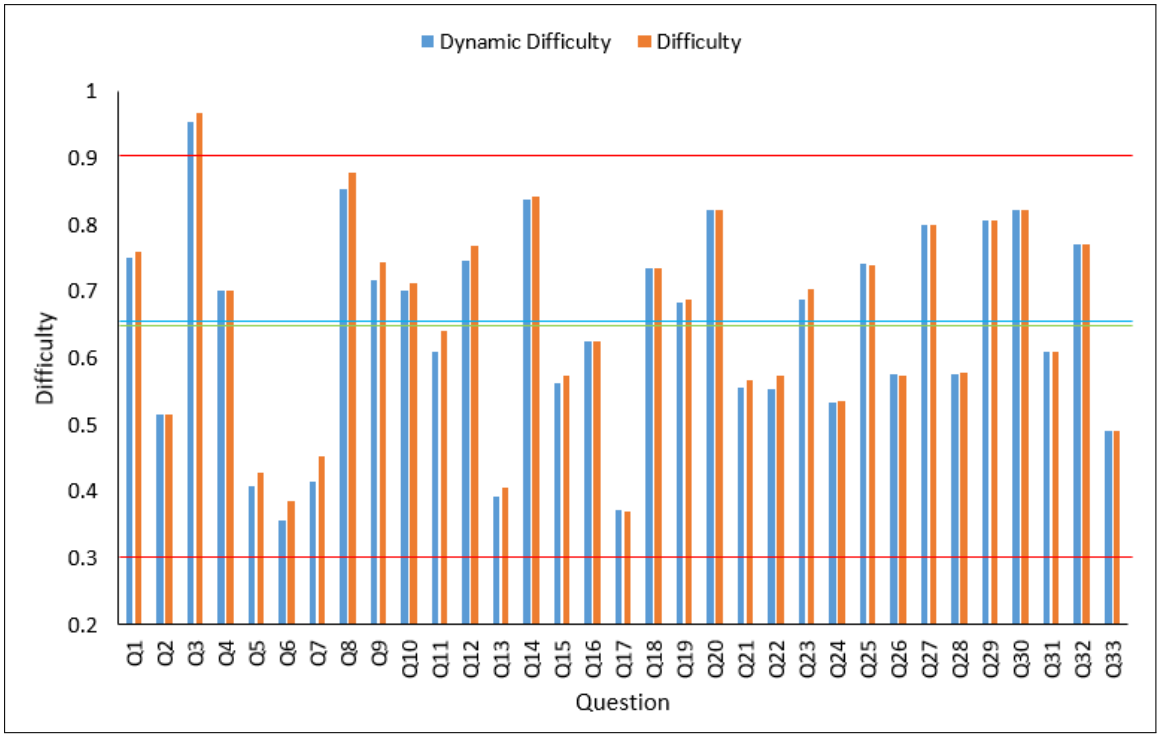


Figure 6.5: Graph showing the dynamic difficulty (blue) and difficulty (orange) of each question on Version 1 of the AMS. The red horizontal lines indicate the lower and upper bounds of the acceptable range of values for the difficulty; the blue horizontal line indicates the mean value of the difficulty; and the green horizontal line indicates the mean value of the dynamic difficulty. Note that higher values indicate *easier* items, whereas lower values indicate *harder* items.

Question	Question type	Dynamic Difficulty	Difficulty
Q1	FRQ	0.75	0.76
Q2	FRQ	0.52	0.52
Q3	FRQ	0.95*	0.97*
Q4	FRQ	0.70	0.70
Q5	FRQ	0.41	0.43
Q6	MRQ	0.36	0.39
Q7	MRQ	0.42	0.45
Q8	MCQ	0.85	0.88
Q9	MCQ	0.72	0.74
Q10	MCQ	0.70	0.71
Q11	FRQ	0.61	0.64
Q12	MCQ	0.75	0.77
Q13	FRQ	0.39	0.41
Q14	MCQ	0.83	0.84
Q15	MCQ	0.56	0.57
Q16	MCQ	0.63	0.63
Q17	FRQ	0.38	0.37
Q18	MCQ	0.74	0.74
Q19	FRQ	0.68	0.69
Q20	FRQ	0.82	0.82
Q21	MRQ	0.56	0.57
Q22	FRQ	0.55	0.57
Q23	FRQ	0.69	0.70
Q24	MCQ	0.53	0.54
Q25	FRQ	0.75	0.74
Q26	MCQ	0.58	0.57
Q27	FRQ	0.80	0.80
Q28	MCQ	0.57	0.58
Q29	FRQ	0.81	0.81
Q30	FRQ	0.82	0.82
Q31	FRQ	0.61	0.61
Q32	FRQ	0.77	0.77
Q33	MRQ	0.49	0.49

Table 6.1: Table showing the dynamic difficulty and difficulty of each question on Version 1 of the AMS. Note that values marked with an asterisk were identified as being problematic, and this convention is applied throughout this thesis.

Table 6.1 and Figure 6.5 show that that dynamic difficulty was larger than the difficulty for Q17, Q25 and Q26. This means that the total number of respondents who attempted these questions found them easier on the whole than the number of respondents who attempted all of the questions. The dynamic difficulty was equal to the difficulty for Q2, Q4, Q16, Q18, Q20, Q27 and Q29, meaning that these questions were of the same difficulty for test-takers who submitted partially complete attempts, and for test-takers who submitted complete attempts. Further, the dynamic difficulty

was also equal to the difficulty for Q30, Q31, Q32 and Q33; this occurred because in these cases, the total number of respondents who attempted these questions was equal to the total number of respondents who attempted all of the questions. The dynamic difficulty was smaller than the difficulty for the other questions, so the total number of respondents who attempted these questions found them harder on the whole than the number of respondents who attempted all of the questions.

In general, the dynamic difficulty is expected to be less than or equal to the difficulty. This is because respondents of lower abilities may be more likely to give up on the test at some point during it, thus not submitting a complete attempt. This trend was observed in the majority of the questions. In the cases where the opposite trend is observed, further explanation was required.

Cases where the dynamic difficulty was greater than the difficulty

In the cases of Q25 and Q26, the number of test-takers who answered each of these questions was 255, whereas the number of test-takers who answered all of the questions was 254. As a result, it would be expected that the values of the dynamic difficulty and difficulty in these two questions would be slightly different because of the extra test-taker's score contributing a small amount to the dynamic difficulty through a stochastic effect. In contrast, the number of test-takers who answered Q17 was 276, which was 22 test-takers more than those who answered all of the questions. As a result, the above explanation based upon a stochastic effect could not be applied to explain why the dynamic difficulty was greater than the difficulty on Q17.

Q17 of Version 1 of the AMS was adapted from Q15 of the original FCI, and it tested understanding of Newton's Third Law. The AMS version of the question is shown in Figure 6.6 below. Q15 of the FCI is known in the literature to be a difficult question (Poutot and Blandin, 2015), and this difficulty could have transferred to the AMS version of the question. As a result, even the most able test-takers may not be expected to get Q17 of Version 1 of the AMS right. For Q17, the values of difficulty and dynamic difficulty remain within the acceptable range of values, but the small fluctuation between the two might indicate a more frequent resort to guesswork than usual on this question.

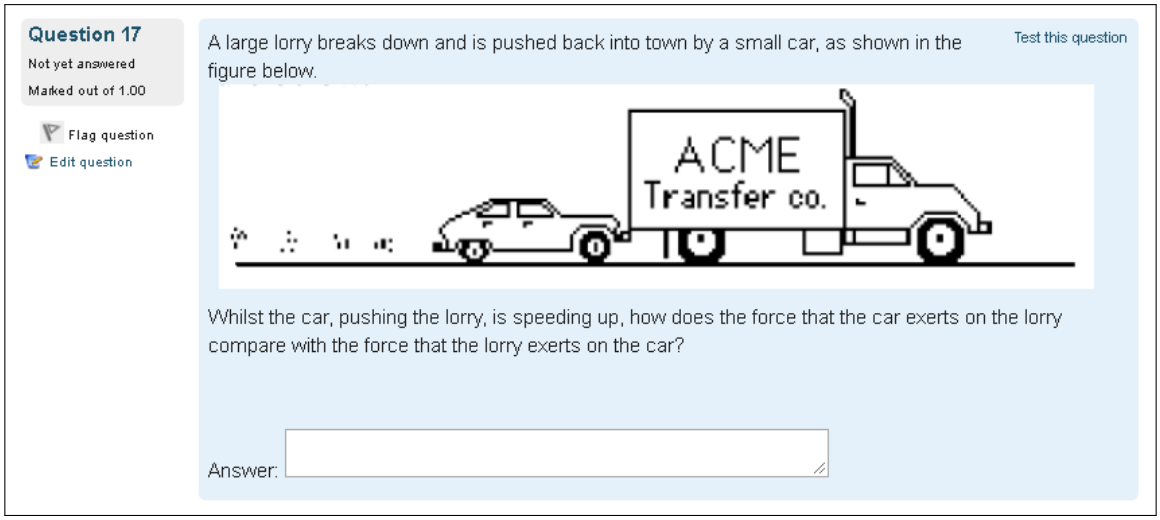


Figure 6.6: Q17 of Version 1 of the AMS, which was adapted from Q15 of the FCI.

In each of the cases where the dynamic difficulty was larger than the difficulty, the difference was never greater than 0.01; this meant that the effect was small, and within what might be expected from random fluctuations where there is some guesswork in the responses, as was explained above. Other cases to consider are those where the question had a difficulty that was out of the acceptable range of values, or close to the boundaries of this acceptable range. As mentioned previously, the acceptable range of values for difficulty are [0.3, 0.9]. One question on the AMS had a difficulty value above 0.9, and five other questions had values that were close to the cut-offs. These are discussed below.

Cases where the difficulty values were high

Q3 had difficulty and dynamic difficulty values that were above the 0.9 cut-off, meaning that almost every test-taker who attempted the questions got it right. This was an essentially new question based on the situation from Q3 of the original FCI, although Q4 of the AMS bears more resemblance to Q3 in the original FCI. Q3 of the AMS is a free-response question, asking the test-taker to identify the force or forces acting on a stone after it is dropped from the roof of a building, and also explicitly instructs test-takers to ignore the effects of air resistance. Examination of the answers showed no flaw in the marking scheme; most test-takers simply answered this question correctly. The very high difficulty value singles Q3 out as a possibly problematic item, with revisions or removal possibly being necessary.

Q8 and Q14 both had difficulty and dynamic difficulty values that were above 0.8, which meant that these questions were two of the easier questions on the AMS. Q8 was adapted from Q6 of the original FCI. It is a multiple-choice question, and it asks the test-taker to identify the trajectory of a marble once it exits a curved channel. Q14 was adapted from Q12 of the original FCI. It is also a multiple-choice question, and it requires the test-taker to identify the trajectory of a cannon ball after it has been fired out of a cannon at the top of a cliff. In both questions, most of the distractor options were rarely selected by the students, with the most frequently selected answers being either the correct answer or one specific incorrect distractor answer. Q8 and Q14 were taken from the original FCI, meaning that they have previously been tested and validated. However, the findings here indicate that the functionality of some of the distractors lead to a potential weakness in these two questions.

From above, most of the distractors on Q8 of the AMS were found to be ineffective, as the majority of the test-takers selected either the correct answer or one other incorrect option. Yasuda et al. (2018) found that students gave correct answers to the FCI version of this question (Q7 of the FCI) by using incorrect lines of reasoning. Furthermore, Traxler et al. (2018) found that the question was biased in favour of males, and even suggested removing it from the original FCI. Similar patterns were identified on Q14 of Version 1 of the AMS, which is another trajectory-based question. This raises questions about what is required to develop effective distractors, particularly in questions that are based on trajectories, since there may only be one viable misconception to base a distractor trajectory path on. However, it is difficult to develop free-response versions of these questions because of the level of description required to specify a path in words. An alternative approach could be to allow students to sketch a trajectory, and to mark the answer based on how close the sketched path is to the desired correct path. Others are investigating the automatic marking of sketches and it has been suggested that this might be incorporated into a version of the FCI (Martinez and Perez, 2010; Martinez, 2020). Combining this approach with the AMS is a possible avenue for future work, but it is beyond the scope of the present study.

Cases where the difficulty values were low

Q6 was the hardest question on the AMS in terms of the dynamic difficulty statistic. Q6 was adapted from Q5 of the original FCI; it is a multiple-response question, and asks the test-taker to identify the forces acting on a marble while it is travelling inside

a curved track. In the usability laboratory testing covered in **Chapter 5**, this question caused problems for some of the participants, as they misinterpreted the diagram by failing to recognize that the track is *flat* on the table. Because the subject backgrounds of the usability testing participants were similar to those of the Version 1 cohort, it is likely that test-takers also had this issue in the large-scale administration of the AMS, leading to the low value for the question difficulty. Similarly to Q8 and Q14, Q6 is taken from the original FCI, so it has previously been tested and validated, but this alone does not mean that it should not be revised. However, rewording the question to encourage students to interpret it in the desired way may be ineffective or counter-productive, since a diagram already accompanies the question to facilitate with its interpretation.

Q17 was the hardest question on the AMS in terms of the difficulty statistic. It is a free-response question adapted from Q15 of the FCI, and it requires the test-taker to apply Newton's Third Law to identify that two forces acting on a car and a truck are equal. As previously noted, Q15 of the FCI is known in the literature as being a difficult question (Poutot and Blandin, 2015), and the concept of Newton's Third Law is known to be a difficult concept for students to master (as discussed in **Section 3.4**). This question could simply be conceptually demanding for the test-takers, leading to the low values of difficulty. The question itself probably does not need any revisions, since its difficulty is not below the cut-off, and it is useful to have more challenging questions as well as more straightforward questions in order to balance the AMS.

More correct answers were given to Q18 of Version 1 of the AMS than to Q17 of Version 1 of the AMS. Q17 of the AMS corresponds to Q15 of the FCI, and it asks students to compare the force that a car has on a truck, while the car is *speeding up* and pushing the truck. Q18 of the AMS corresponds to Q16 of the FCI, and it asks students to compare the force that the same car has on the same truck, when the car is pushing the truck at a *constant speed*. For Version 1 of the AMS, Q17 was a free-response question, whereas Q18 was a multiple-choice question, and it is possible that students in general found the multiple-choice variant of the question to be easier. However, it is important to note that while the questions are based on the same situation, they are not identical, and it is possible that other factors cause Q17 to be answered better than Q18.

Rebello and Zollman (2004) noted that FCI Q16 makes use of the wording *"constant cruising speed"*, which is not present in the preceding FCI Q15. It is possible that this wording guides students to the correct answer by using a faulty line of reasoning, and this idea is supported by Yasuda et al. (2018) and Galloway (2019) who found that students had linked the idea of moving at *constant speed* to the forces being equal. It is possible that students in the Version 1 cohort used such reasoning to answer Q18 correctly, although it is not possible to verify this with Version 1 student responses, since Q18 was a multiple-choice question in Version 1 of the AMS. However, Q18 was a free-response question in Version 2 of the AMS (the Version 2 study is covered in **Chapter 7**), and going through these responses revealed that a small number of students did explicitly make use of the incorrect *constant speed* line of reasoning, invoking Newton's First Law, to give a correct answer to this question.

Summary

Overall, 32 out of the 33 questions on the AMS had a difficulty value and dynamic difficulty value that were in the acceptable range of [0.3, 0.9]. The mean value of the difficulties of the individual questions was 0.65, and the mean value of the dynamic difficulties of the individual questions was 0.65. Both of these values were within the acceptable range for difficulty, which implied that the AMS was functioning in the desired way in terms of its difficulty.

6.3.3 Discrimination and point biserial coefficient

The discrimination and point biserial coefficient values were calculated for the Version 1 AMS responses marked by the UHM, and the results are given in Table 6.2 and Figure 6.7 below.

Question	Question Type	Discrimination	Point Biserial Coefficient
Q1	FRQ	0.50	0.51
Q2	FRQ	0.46	0.42
Q3	FRQ	0.06*	0.22
Q4	FRQ	0.39	0.37
Q5	FRQ	0.67	0.56
Q6	MRQ	0.75	0.64
Q7	MRQ	0.78	0.65
Q8	MCQ	0.26*	0.42
Q9	MCQ	0.47	0.55
Q10	MCQ	0.51	0.49
Q11	FRQ	0.53	0.50
Q12	MCQ	0.50	0.56
Q13	FRQ	0.76	0.66
Q14	MCQ	0.38	0.47
Q15	MCQ	0.81	0.78
Q16	MCQ	0.65	0.63
Q17	FRQ	0.49	0.45
Q18	MCQ	0.43	0.43
Q19	FRQ	0.58	0.58
Q20	FRQ	0.39	0.46
Q21	MRQ	0.81	0.72
Q22	FRQ	0.58	0.56
Q23	FRQ	0.29*	0.36
Q24	MCQ	0.58	0.44
Q25	FRQ	0.42	0.45
Q26	MCQ	0.67	0.62
Q27	FRQ	0.42	0.50
Q28	MCQ	0.78	0.65
Q29	FRQ	0.24*	0.24
Q30	FRQ	0.32	0.34
Q31	FRQ	0.74	0.62
Q32	FRQ	0.40	0.43
Q33	MRQ	0.82	0.74

Table 6.2: Table showing the discrimination and point biserial coefficient of each question on Version 1 of the AMS.

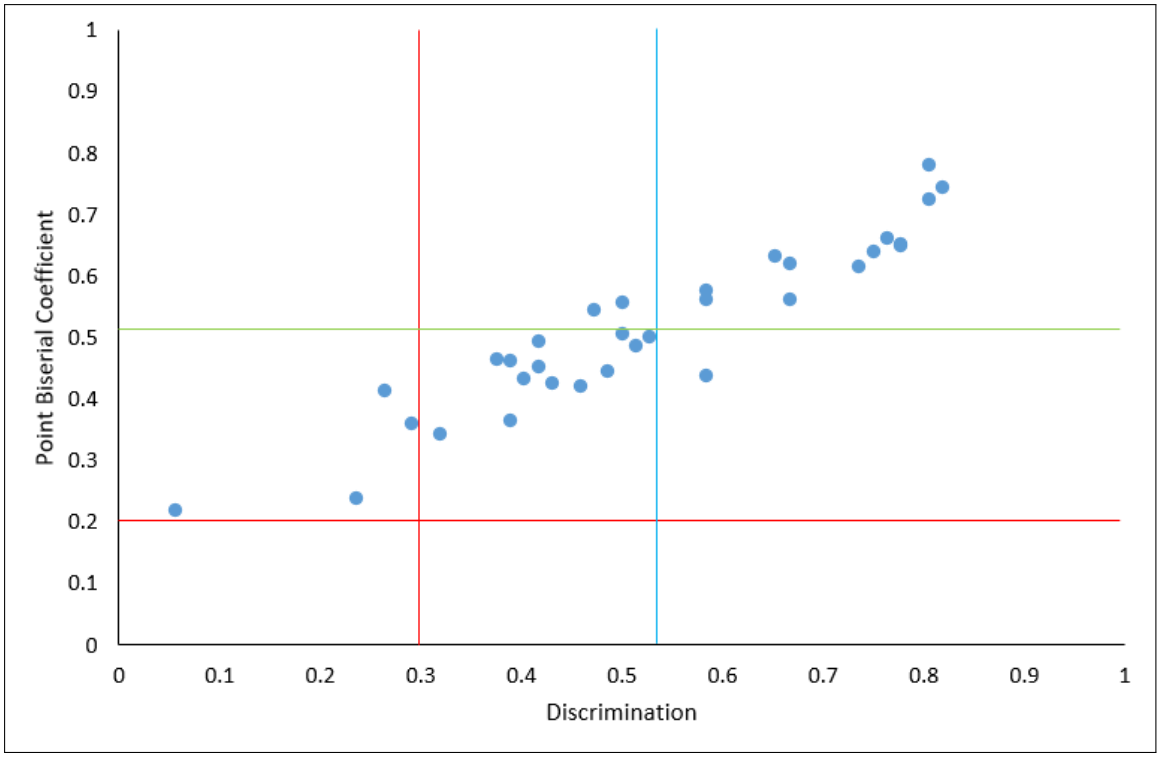


Figure 6.7: Graph showing the point biserial coefficient and the discrimination of each question on Version 1 of the AMS. The red horizontal line represents the lower bound of the acceptable values for point biserial coefficient, and the green horizontal line represents the mean value of the point biserial coefficient. The red vertical line represents the lower bound of the acceptable values for discrimination, and the blue vertical line represents the mean value of the discrimination.

The acceptable range of values for discrimination are [0.3, 1], and the acceptable range of values for point biserial coefficient are [0.2, 1]. From Table 6.2 and Figure 6.7, four questions had discrimination values that were outside the acceptable range; of these, two of the questions also had lower values for the point biserial coefficient. In contrast, three questions had values for discrimination and point biserial coefficient that were at the higher end of the acceptable range of values. These cases are discussed below.

Cases where the discrimination and/or point biserial coefficient were low

Q3 had a discrimination value that was outside the acceptable range of values, whereas its point biserial coefficient was at the lower end of the acceptable range of values. The features and workings of Q3 were previously identified as problematic due to its difficulty values also being outside the acceptable range. The item had almost no discriminatory power, and this is consistent with the fact that almost all test-takers, regardless of ability, got this question right. Additionally, the item had

lower alignment with the rest of the test, which could be a consequence of Q3 being a new question added to the already well-established FCI. Several issues raised with Q3 through calculation of its difficulty, discrimination and point biserial coefficient implied that the question required re-wording or removal to resolve the issue.

Q8 had an acceptable value for the point biserial coefficient, but its discrimination value was outside the acceptable range of values. Q8 was previously found to be too easy in the difficulty aspect of this analysis, which is consistent with its low discriminatory power since test-takers of all abilities are able to get this question right.

Q23 had acceptable values for both point biserial coefficient and difficulty, so it was not a problematic item in these respects. In contrast, Q23 had a discrimination value which was slightly below the acceptable range of values. Q23 was the question in which 20 participants gave up their attempt on Version 1 of the AMS; it was previously postulated in **Subsection 6.3.1** that test-takers with a lower previous exposure to physics gave up after answering this question, which would instead indicate that Q23 would have high discriminatory power. However, the discrimination statistic is calculated using complete AMS attempts only, which means that the scores of the 20 students who gave up on their attempts were not taken into account for the calculation of the discrimination value of Q23. As a result, since the students of lower abilities were not included in this calculation, the discrimination value for Q23 would be expected to be lower by definition, because there was a smaller range of abilities to discriminate between *ab initio*.

Q29 had a lower (but still acceptable) point biserial coefficient value, but its discrimination value was below the acceptable range of values. It is a free-response question, and it asks the test-taker to identify what happens to the speed of a box when the force being exerted on it is doubled. Q29 is not a new question added to the AMS, since it was adapted from Q26 of the original FCI; it follows that Q29 would be expected to align well with the rest of the AMS. The correct answer is that the speed *increases*, although the speed does *not double*, since force and velocity are not related in this way. As a result, answers that state that the speed does double need to be marked as incorrect. Q29 did not have any issues with difficulty, as its difficulty values were within the acceptable range. In addition, the wording of Q29 was also not likely to be an issue, since it was adapted from its original FCI counterpart, which had already been tested and validated. As a result, the discrimination issues with Q29

may instead arise from a factor which cannot be measured quantitatively with CTT statistics. For example, the context of the situation presented in the question may be difficult for students from particular demographic groups to interpret.

Cases where the discrimination and/or point biserial coefficient were high

Q15 had the highest value for the point biserial coefficient out of all of the questions as well as having a high discrimination value, positioning it in the top right-hand corner of Figure 6.7. Q15 was adapted from Q13 of the original FCI. It is a multiple-choice question, and it asks the test-taker to identify what forces act on a ball after it had been thrown upwards. The effective performance of Q15 is in stark contrast to the ineffective performance of Q3, which is similar in conceptual content. One important difference between Q15 and Q3 is that Q15 is multiple-choice (and very similar to Q13 in the original FCI), whereas Q3 is free-response, and this may have been a factor which contributed to the differences in the CTT statistics between the two questions. However, the situation in Q15 involves the ball being tossed upwards, whereas the situation in Q3 involves the ball being dropped; this makes Q15 more conceptually demanding than Q3, as it requires test-takers to do more than simply identify the name of a force. This offered a possible explanation for the differences in discriminatory power of the two items, since only the higher-performing students would be expected to answer conceptually more demanding questions well. In addition, the fact that Q15 was taken directly from the FCI offers a logical explanation to the differences in alignment to the rest of the test between the two questions, because Q3 was not originally an FCI question.

Q33 had the highest value for discrimination out of all of the questions in addition to having a high point biserial coefficient value. Q33 was adapted from Q30 of the original FCI. It is a multiple-response question, and it asks the test-taker to identify what forces are acting on a tennis ball while it is in flight. Q33 is different from the other questions on the AMS because it asks the test-taker to take into account forces due to the air, whereas air resistance is supposed to be ignored for the other questions. By looking through the responses given to this question, the high discrimination was a reflection of the highest-scoring students being those who realized that there is no residual force acting on the tennis ball from being hit. It is also possible that since this was the last question on the AMS, less keen students stopped paying attention to this question and gave incorrect answers to it as a result, contributing the the observed

effect. The high point biserial coefficient could be a reflection of the content of the question being highly consistent with the rest of the AMS.

Q21 was the third point in the top right-hand corner of Figure 6.7, showing that it had high values for both the discrimination and the point biserial coefficient. Q21 is a multiple-response question adapted from Q18 of the FCI. The question asks the test-taker to identify the forces acting on a boy on a swing while he is in motion. A few features of this question could explain the high discriminatory capabilities of this question. First, as this is a multiple-response question, it requires students to identify more than one force in order to give the correct answer, which is a more conceptually demanding task that only the higher-scoring students may be expected to complete. Second, the question features a diagram, and this may help medium-level students to visualize the situation, and scaffold them towards the correct answer (Dawkins et al, 2017). However, the diagram may not offer much assistance to higher or lower achieving students; this is because higher-achieving students do not require the scaffolding in order to provide the correct answer, whereas lower-achieving students are not able to use the provided scaffolding to reach the correct answer. As was the case for Q15 and Q33, the high value for the point biserial coefficient might be expected since Q21 was adapted from an FCI question.

Summary

Overall, 29 questions on the AMS had discrimination values that were within the acceptable range of values. This meant that on the question level, 29 out of the 33 questions could differentiate between higher and lower performing students. For the test level statistics, the mean value of the discrimination of the individual questions was 0.53, which was also within the acceptable range for the discrimination value. Taking this together with the question-level findings, this implied that the overall AMS was capable of distinguishing between the higher-performing and lower-performing students.

For the point biserial coefficient, all 33 questions had values that were within the acceptable range of values; practically, this means that on the question level, every question on the AMS aligned to test similar concepts. For the test level statistics, the mean value of the point biserial coefficients of the individual questions was 0.52, and this was also within the acceptable range for the point biserial coefficient value.

Taking this together with the question-level findings, this implied that the overall AMS contained questions that assessed similar topics. These results taken together with those for the difficulty and discrimination statistics provided important evidence for the overall functionality of the AMS questions.

6.3.4 Overall functioning

For consideration of the overall test function, the mean values of the difficulty, discrimination and point biserial coefficient were calculated. These are shown in Table 6.3 below, together with the Kuder-Richardson reliability and the Ferguson's delta values for the entire test.

Classical Test Theory Statistic	Value	Desired values
Mean difficulty	0.65	[0.3, 0.9]
Mean discrimination	0.53	≥ 0.3
Point biserial coefficient	0.52	≥ 0.2
Kuder-Richardson reliability	0.92	≥ 0.7
Ferguson's delta	0.98	≥ 0.9

Table 6.3: Table showing the overall CTT statistics for Version 1 of the AMS.

Using the previously calculated difficulty values for each of the items, the standard deviation of the total scores, and $K = 33$ for the 33 AMS items, a Kuder-Richardson reliability of 0.92 was calculated for Version 1 of the AMS. This was above the threshold value of 0.7, and this showed that the AMS was reliable overall. Next, taking the values of the frequency for each possible score, $N = 254$ as the number of test-takers, and $K = 33$ for the number of items, a Ferguson's delta value of 0.98 was found for the AMS. This was above the acceptable value of 0.9, and this showed that the overall AMS could discriminate between lower and higher scoring students. This δ calculation agreed with what was concluded from the analysis of the discrimination coefficient on each of the questions previously.

The mean difficulty was within the acceptable range of values, showing that the AMS was not too easy nor too hard for the test-takers. The mean discrimination was within the acceptable range of values, and the Ferguson's delta value was also within the acceptable range of values; this showed that the AMS was able to distinguish between test-takers of higher performance and lower performance. The mean point biserial coefficient value was within the acceptable range, demonstrating that the items

on the AMS tested similar content. Further, the Kuder-Richardson reliability was within the acceptable range of values, illustrating that the AMS was reliable. Taken together, these results showed that the AMS questions were generally functioning at an acceptable level. However, possible issues were raised by the small number of questions with statistics that were outside of the acceptable range of values. Possible improvements were suggested for these items, noting that one of the objectives of the overall study is to ensure that the AMS questions are functioning at the highest possible level.

6.4 Results and Discussion: AMS Version 1 IRR study

6.4.1 Marking agreement and Cohen's kappa

The marking agreement and Cohen's kappa values for the Version 1 UHM against the Version 1 computer marking were calculated. The results are given in Table 6.4 below, together with data pertaining to the number of times the UHM disagreed with the computer marker and the nature of these disagreements. Note that a *false positive* refers to instances where the computer marked the answer as *correct* and the UHM marked the answer as *incorrect*; whereas a *false negative* refers to instances where the UHM marked the answer as *correct* and the computer marked the answer as *incorrect*. In addition, the table contains only free-response questions, since the IRR calculations only give meaningful results where the marking is subjective. The overall trends emerging from the results are summarized below the table.

Question	Number of responses	Number of disagree-ments	Number of false positives	Number of false negatives	Marking agree-ment	Cohen's kappa
Q1	328	10	1	9	0.97	0.92
Q2	307	33	7	26	0.89*	0.79*
Q3	305	12	10	2	0.96	0.38*
Q4	304	78	62	16	0.74*	0.28*
Q5	301	10	4	6	0.97	0.93
Q11	280	31	21	10	0.89*	0.76*
Q13	277	12	6	6	0.96	0.91
Q17	276	6	5	1	0.98	0.95
Q19	275	22	19	3	0.92*	0.81
Q20	275	27	4	23	0.90*	0.71*
Q22	275	24	2	22	0.91*	0.83
Q23	275	32	22	10	0.88*	0.72*
Q25	255	6	6	0	0.98	0.94
Q27	255	9	6	3	0.96	0.89
Q29	255	15	10	5	0.94*	0.80
Q30	254	44	24	20	0.83*	0.38*
Q31	254	2	1	1	0.99	0.98
Q32	254	24	12	12	0.91*	0.73*

Table 6.4: Table showing the number of times the UHM disagreed with the Version 1 computer marking on the Version 1 free-response AMS questions and the nature of these disagreements, as well as the corresponding marking agreement and Cohen's kappa values for the UHM against the Version 1 computer marking.

The acceptable range of values for marking agreement are [0.95, 1], whereas the acceptable range of values for Cohen's kappa are [0.8, 1]. From Table 6.4, seven questions had acceptable values for both marking agreement and Cohen's kappa; three questions had an acceptable value for Cohen's kappa, but a lower value for marking agreement; one question had an acceptable value for marking agreement, but a lower value for Cohen's kappa; and seven questions had lower values for both marking agreement and Cohen's kappa. Each of these scenarios is discussed below.

Cases with acceptable values for both marking agreement and Cohen's kappa

For Q1, Q5, Q13, Q17, Q25, Q27 and Q31, the values for the marking agreement and Cohen's kappa were acceptable, meaning that the marking rules were functioning well in all of these questions. However, various false positive and false negative cases were encountered in each of these questions.

There were two main types of false positive cases. The first type of false positive arose when the computer recognized an incorrect answer as an acceptable misspelling of a correct answer (for example, accepting *there* as a misspelling of *three*), and these cases were handled by altering the marking rules to differentiate between the different spellings. The second type of false positive occurred when the computer marked answers that contained both correct and incorrect aspects. In most of these cases, it was possible to resolve the false positives by developing marking rules to negate on the incorrect parts of the answers. However, this approach was not effective in some cases, because changing the marking rules in this way had an adverse effect on the overall computer marking by adding new false negative cases, an effect previously alluded to in the work of Butcher and Jordan (2010).

In contrast, there was only one type of false negative case; this arose when students gave correct answers that were not recognized by the computer marking rules. In many of the cases, it was possible to cover the false negatives by adding extra marking rules to cover the alternative correct answers. In some cases, the approach was not employed because the wordings used in the answers were too specific, meaning that developing extra rules to account for them would have been tantamount to *over-fitting* (Zehner et al, 2016).

Cases with acceptable values for Cohen's kappa, but lower values for marking agreement

In the cases of Q19, Q22, and Q29, the questions had acceptable Cohen's kappa values, but the corresponding marking agreement values were slightly below the acceptable value. The false positive and false negative cases were of the same types as those outlined above, although they were more numerous for these questions. In spite of this, there were not any serious concerns about the functionality of the marking rules for these questions, since the high Cohen's kappa values indicated that agreement between the UHM and the computer marking was not the result of random chance. This outcome can be explained as follows. As mentioned previously, marking agreement does not take into account chance agreement, so it is an overestimate. The acceptable threshold for marking agreement was set high by Butcher and Jordan (2010), but this threshold was set empirically. Therefore, if the Cohen's kappa value is above the threshold, then it is likely that the marking rules are functioning in the expected way, although this needs to be monitored carefully on a case-by-case basis. In addition, no systematic problems were uncovered when working back through the false positive and false negative cases to improve the effectiveness of the marking rules on questions Q19, Q22 and Q29. However, further testing was required to check that changes made to the marking rules had the desired effect.

Cases with acceptable values for marking agreement, but lower values for Cohen's kappa

Q3 had an acceptable value for the marking agreement, but the Cohen's kappa was outside the acceptable range of values. The low value of Cohen's kappa for this question indicated that the high level of agreement between UHM and computer here may have arisen out of random chance. Since almost everyone who attempted Q3 got it right, the data is skewed in favour of the *correct* classification category for this question. The resulting effect observed is an example of the counter-intuitive result that for skewed data, coders can agree on the classification of the majority of the items, but the value of advanced IRR coefficients such as Cohen's kappa can remain low (Artstein and Poesio, 2008). As a result, the low kappa value was an unavoidable consequence of the problematic high difficulty value of this question.

Cases with lower values for marking agreement and Cohen's kappa

Q2, Q4, Q20, Q23, and Q32 each had values for both the marking agreement and Cohen's kappa that were outside the acceptable range of values, meaning that the Version 1 marking rules were under-performing on each of these questions. The false positive and false negative cases on these questions were the same types as those outlined earlier in **Subsection 6.4.1**. As a result, the strategy of countering false positive answers by adding marking rules to negate on incorrect aspects of the answers was found to be effective in dealing with most of the false positive cases within these questions. In addition, the approach of adding marking rules to cover false negative answers was capable of handling most of the false negative cases in these questions. However as was the case previously, new responses were required to test whether the changes to these marking rules had improved the functionality of the computer marking.

Q11 had values for both the marking agreement and Cohen's kappa that were outside the acceptable range of values, suggesting that there were problems with the Version 1 marking rules. There were 31 cases where the computer marking did not concur with the UHM; 21 of these were false positives, and 10 were false negatives. The large number of false positive cases here indicated that the computer mark scheme was not matching up well to the intended correct answer. The question requires students to compare the speed of a puck before and after it has been kicked. It is *part 2* of a four-part question that runs from Q10 to Q13, and it shown in Figure 6.8 below.

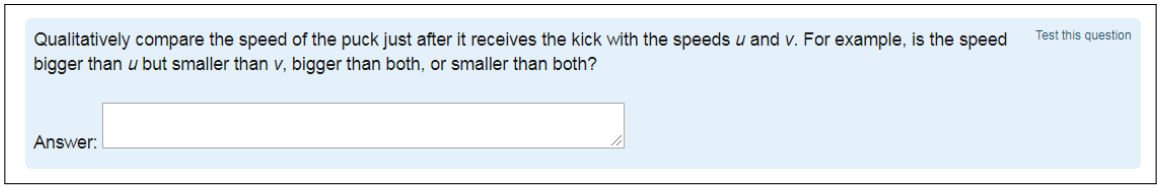


Figure 6.8: Q11 of Version 1 of the AMS, which is adapted from Q9 of the FCI.

The correct answer sought is one that identifies that the speed after the kick will be *bigger than both* of the initial speeds u and v. However, many answers to the question indicates that students had misunderstood the question, as they compared the speeds u and v themselves:

"V is greater than u"

"v bigger than u"

Or by simply pointing out that they didn't understand the question:

"Do not understand the question u or v"

These misinterpretations were not well picked up by the marking rules. In addition, attempts to improve the marking rules by adding extra rules to account for correct and incorrect answers were ineffectual, since the addition of new rules caused alternative false positive and false negative cases to arise. Additional responses and further work were required to resolve this issue.

Q30 had values for both the marking agreement and Cohen's kappa that were outside the respective acceptable range of values, again implying deficiencies in the Version 1 marking rules. There were 44 cases where the UHM and computer marking did not agree; 24 of these were false positives, and 20 were false negatives. The high number of false positive and false negative cases arose because the intended marking scheme did not translate well to computer marking. The question required students to identify what would happen to the speed of a box after a woman stops pushing it. It is *part 3* of a three-part question that runs from Q28 to Q30, and it is shown in Figure 6.9 below.

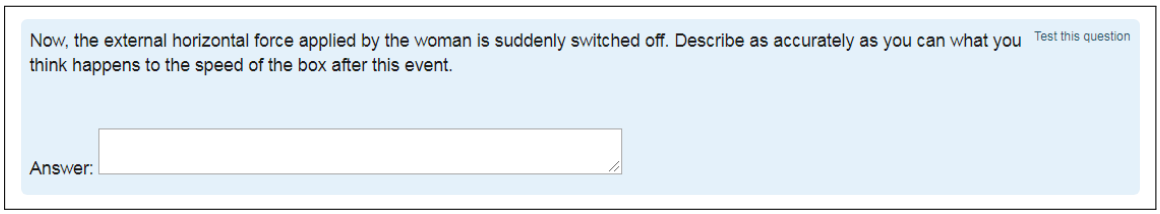


Figure 6.9: Q30 of Version 1 of the AMS, which is adapted from Q27 of the FCI.

The correct answer sought is one that identifies that the box *slows down*. Answers that identify that the box *slows down and stops* are also correct, but answers that only state that the *box will stop* were deemed to be incorrect because, according to Newton's laws of motion, the slowing effect is not instantaneous. This marking metric did not translate well to the marking rules, for example the following answers were marked as correct by the computer, but as incorrect by the UHM.:

"it will stop"

"it decreases immediately and stops"

As was the case with Q11, attempts to improve the marking rules by adding rules based on the false positive and false negative cases were not effective, since the addition of new rules caused alternative problem cases to arise. These findings indicated that automatic marking may not be viable for a free-response version of Q30; however, further responses and testing were required to test this claim.

6.4.2 Back-testing the Version 2 marking rules against the Version 1 responses

The changes to the marking rules detailed above were used to develop the Version 1 AMS computer marking rules into the Version 2 AMS computer marking rules. The Version 2 rules were then back-tested against the Version 1 responses that were used to build them, as a check for consistency. The results of this back-testing are given in Table 6.5 below. Note that the Version 2 AMS computer marking rules were subsequently used to mark AMS responses gathered in the academic year 2018-2019; this use is detailed in **Chapter 7**.

Question	Number of responses	Number of disagreements	Number of false positives	Number of false negatives	Marking agreement	Cohen's kappa
Q1	328	1	0	1	1.00	0.99
Q2	307	0	0	0	1.00	1.00
Q3	305	9	1	8	0.97	0.73*
Q4	304	11	7	4	0.96	0.91
Q5	301	3	3	0	0.99	0.98
Q11	280	30	21	9	0.89*	0.77*
Q13	277	9	5	4	0.97	0.93
Q17	276	5	4	1	0.98	0.96
Q19	275	12	7	5	0.96	0.90
Q20	275	3	3	0	0.99	0.96
Q22	275	4	2	2	0.99	0.97
Q23	275	7	2	5	0.97	0.94
Q25	255	0	0	0	1.00	1.00
Q27	255	3	2	1	0.99	0.96
Q29	255	2	0	2	0.99	0.98
Q30	254	41	28	13	0.84*	0.36*
Q31	254	0	0	0	1.00	1.00
Q32	254	7	4	3	0.97	0.92

Table 6.5: Table showing the number of times the UHM disagreed with the Version 2 computer marking on the Version 1 free-response AMS questions and the nature of these disagreements, as well as the corresponding marking agreement and Cohen's kappa values for the UHM against the Version 2 computer marking.

Three different cases emerged when considering the IRR statistics and improving the marking rules. In the first case, both the marking agreement and Cohen's kappa values were within the respective acceptable ranges, and the marking rules did not require much modification as a result. Q1, Q5, Q13, Q17, Q25, Q27 and Q31 were the questions in this scenario, and the IRR results from these questions were encouraging for the development of automated marking schemes. Out of these questions, Q5, Q17 and Q31 all tested the concept of Newton's Third Law, which may suggest that it is easier to author automated marking schemes for questions based on Newton's Third Law than for other concepts tested in the FCI.

In the second case, one or both of the marking agreement and Cohen's kappa values were below the respective accepted values. However, these questions had problems that could be resolved by considering the false positive and false negative answers, and developing the marking rules further to handle these cases. Q2, Q3, Q4, Q19, Q20, Q22, Q23, Q29 and Q32 were the questions in this scenario, and the corresponding IRR statistics from the back-testing showed improvement for these questions. However, further responses and IRR testing was required to check that the rules also worked on other responses.

The cases of Q11 and Q30 were not straightforward to deal with, and this pair of questions made up the third scenario. For both questions, the marking agreement and Cohen's kappa statistics had values that were outside the acceptable range, which highlighted a need to look again at the marking rules. However, unlike the other cases, the strategy of using the false positives and false negatives to improve the marking rules was found to be ineffective; this was because of the questions themselves, and the sorts of answers that they drew from students. More student responses were required to further develop the marking rules in these cases, with the possibility of reverting one or both of the questions to multiple-choice format, or of rewording the questions, if progress could not be made.

6.4.3 Discussion of the approach used to develop the computer marking rules

In the Version 1 IRR study, the effectiveness of the computer marking was tested by comparing it to human marking, and by seeing how many cases arose where the computer and human marking disagreed. The general strategy adopted to improve the computer marking was to use the false positive cases to develop new marking rules to mark answers containing incorrect information as incorrect; and to use the false negative cases to add extra rules that allowed for correct answers that were not previously covered by the marking rules to be marked as correct. The strategy was found to be mostly effective in this study; however, potential issues relating to the use of false positives and false negatives to develop new marking rules have been identified.

For the use of false positives to create new rules to mark answers as incorrect, there exists the possibility of correct answers being caught by the system and being erroneously marked as incorrect (Butcher and Jordan, 2010). This can lead to the creation of new false negative cases, with a balance being difficult to obtain between negating false positives and including false negatives. In contrast, different concerns arise when using false negatives to develop new marking rules to account for a wider variety of correct answers. Because of the open-ended format, there are many ways in which students can answer any particular free-response question; as a result, it can be difficult to develop marking rules to cover all of the possible answers (Sychev et al., 2020). In addition, it is possible that the marking rules developed using this process cover answers specific to the data that was used to develop the rules, which raises the possibility of an *over-fitting* concern (Zehner et al., 2016). Gathering additional responses to further develop and test the marking rules was required to address the concerns with the rule development process outlined above.

The issue of balancing false positive and false negative cases leads onto the question of whether it is worse to have more false positive or false negative cases in the first place. False positives mean that students can get a false sense of their own understanding in some contexts, which could have serious consequences in the medical education context (Ali et al., 2016; Mahjabeen et al., 2017). The consequences in the formative Physics Education Research context could be deemed to be rather less severe. False negatives mean that students who have understanding of the topic do not get the credit that they deserve for answering the question. This could potentially lower a student's overall

course grade in a summative setting, which is a situation that should be avoided where possible. In the context of the current work, it is therefore reasonable to propose that reducing the number of false negatives would gain priority over reducing the number of false positives. For concept inventories it should be noted that false negatives and false positives can both result in an educator having a false impression of their student cohort's understanding, or of the understanding of an individual student.

Different questions are more likely to produce more false positives or more false negatives (Sychev et al., 2020), so the solution to this problem may lie at the question-setting level rather than at the automated marking level. This agrees with the previous observation of Butcher and Jordan (2010) that computer marked assessment can be made more effective by changing question wording, as well as by modifying the automated marking schemes. Beyond the Physics Education Research context, different subject disciplines have vastly different types of answers to mark (Sarrouti and El Alaoui, 2020), and different approaches to developing marking rules will also have different levels of effectiveness for other subject areas. One way to improve the accuracy of automated marking in general would be to set questions up to return exact answers (Sarrouti and El Alaoui, 2020), which could be achieved by having students answer questions that request a specific word, or a yes/no response. However, this approach would defeat the objective of the current study, since the aim is to learn about students' understanding and misunderstandings by giving them the freedom to express themselves using their own words.

6.4.4 Findings related to testing the human marking

Question	Marker 1 vs. UHM (MA)	Marker 1 vs. UHM (CK)	Marker 2 vs. UHM (MA)	Marker 2 vs. UHM (CK)	Marker 3 vs. UHM (MA)	Marker 3 vs. UHM (CK)	Marker 4 vs. UHM (MA)	Marker 4 vs. UHM (CK)	Marker 5 vs. UHM (MA)	Marker 5 vs. UHM (CK)
Q1	1.00	1.00	1.00	0.99	0.99	0.98	0.99	0.97	1.00	1.00
Q2	0.97	0.95	1.00	0.99	0.97	0.94	0.98	0.95	0.98	0.96
Q3	0.95	0.57*	0.96	0.65*	0.97	0.69*	0.82*	0.20*	0.97	0.73*
Q4	0.99	0.97	0.98	0.95	0.76*	0.26*	0.98	0.95	0.98	0.94
Q5	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99
Q11	0.99	0.97	0.98	0.96	0.98	0.96	0.98	0.95	0.99	0.99
Q13	0.99	0.98	0.99	0.98	0.97	0.94	0.95	0.89	0.99	0.97
Q17	0.99	0.97	1.00	0.99	0.98	0.95	0.99	0.98	1.00	0.99
Q19	0.79*	0.58*	0.79*	0.59*	0.94*	0.86	0.82*	0.51*	0.91*	0.79*
Q20	0.99	0.96	0.99	0.95	1.00	0.99	0.99	0.95	0.99	0.96
Q22	0.96	0.93	0.95	0.91	0.92*	0.85	0.99	0.98	0.99	0.98
Q23	0.99	0.97	0.98	0.95	0.97	0.94	0.98	0.96	0.99	0.98
Q25	1.00	0.99	1.00	1.00	0.99	0.98	0.98	0.96	1.00	0.99
Q27	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.96	1.00	1.00
Q29	0.98	0.94	1.00	1.00	0.97	0.91	0.99	0.98	0.99	0.96
Q30	0.98	0.95	0.98	0.92	0.89*	0.67*	0.92*	0.73*	0.96	0.86
Q31	0.99	0.98	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.99
Q32	0.98	0.93	0.94*	0.84	0.95	0.87	0.92*	0.74*	0.96	0.90

Table 6.6: Table showing the marking agreement (MA) and Cohen's kappa (CK) values of each human marker against the Version 1 UHM for the free-response AMS questions.

It was also important to check whether the human markers used to build the UHM were consistent. To this end, the IRR statistics were calculated for each of the human markers against the Version 1 UHM. The results are presented in Table 6.6 above. Note that the abbreviation *MA* in Table 6.6 is used to denote *marking agreement*, and the abbreviation *CK* is used to denote *Cohen's Kappa*. There was a high level of agreement between each of the human markers and the UHM on questions Q1, Q2, Q5, Q11, Q13, Q17, Q20, Q23, Q25, Q27, Q29 and Q31, because both the marking agreement and Cohen's kappa values were within the respective acceptable range of values for these 12 questions. This meant that the UHM was highly self-consistent for these questions. The statistics for the other questions are discussed below.

Q3 asked students to identify the forces acting on a stone after it was dropped from the top of a building. For Q3, all five of the Markers had Cohen's kappa values that fell below the acceptable value when compared to the UHM, and Marker 4 additionally had a marking agreement value that was below the acceptable value when compared to the UHM. This indicated that marking Q3 was problematic for the human markers. Apart from the case of Marker 4, the marking agreement between the Markers and the UHM was good, but the corresponding values of Cohen's kappa were not. This can be explained as follows. Almost all of the responses to Q3 were marked as correct by the different markers, which leads to a high level of marking agreement, since most of the responses have been marked as correct. However, Cohen's kappa is designed to account for random agreement, and abnormally high levels of agreement between the markers can count towards this, as was observed in this question. This meant that there was not a fault in the marking guidelines here; rather, the abnormally high percentage of correct answers to Q3 was highlighted by the Cohen's kappa calculation.

Q4 asked students to identify what happens to the speed of a stone after it is dropped from a building. For this question, Markers 1, 2, 4 and 5 had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM; Marker 3 instead had values for the marking agreement and Cohen's kappa that were outside the respective acceptable range of values. This difference arose because Marker 3 had consistently marked answers that referred to *terminal velocity* as correct, whereas the other markers had consistently marked such answers as incorrect.

Q19 required students to identify the forces acting on an elevator as it moved up a smooth shaft. For this question, all five of the markers had values for the marking

agreement that were outside the acceptable range when compared to the UHM. In addition, Markers 1, 2, 4 and 5 had Cohen's kappa values that were outside the acceptable range of values when compared to the UHM. This means that in the case of Q19, the level of agreement between the individual human markers and the UHM was not good. This is a similar situation to that observed for Q3, although the explanation for this effect is more straightforward in the case of Q19. Here, the marking guidelines pertaining to what could be accepted as an acceptable synonym for the *tension* force was ambiguous in the Version 1 marking guidelines, which led to the high levels of disagreement between the different markers. As a result, the Version 2 marking guidelines were improved with this consideration in mind.

Q22 required students to identify a time interval where a pair of moving blocks had the same speed. For Q22, Markers 1, 2, 4 and 5 all had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM; Marker 3 had an acceptable Cohen's kappa value, but their marking agreement value was slightly below the acceptable value. This discrepancy occurred because Marker 3 had been strict on marking student answers to this question as correct only if the answer specified an *interval* rather than a *point*, whereas the other markers accepted answers that specified a *point* rather than an *interval*.

Q30 required students to describe what happens to the motion of a box after the external force being used to push it is removed. For this question, Markers 1, 2 and 5 had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM; whereas Markers 3 and 4 had values for both the marking agreement and Cohen's kappa that were outside the respective acceptable range of values. This discrepancy arose for Marker 3 because they strictly marked anything that implied that the effect of the box stopping was instantaneous as incorrect, whereas the other markers had been more lenient. For Marker 4, the discrepancy occurred instead because their marking was inconsistent, as they sometimes marked answers that implied that the box stopped immediately as correct, and sometimes marked these answers as incorrect.

Q32 asked students to identify the forces acting on an office chair at rest on a floor. Markers 1, 3 and 5 had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM for this question; Marker 2 had an acceptable Cohen's kappa value but their marking agreement value was slightly below the acceptable value; Marker 4 had values for both the marking agreement and Cohen's kappa

that were outside the respective acceptable range of values. In the case of Marker 2, the discrepancy arose because they had been less generous on what constituted an acceptable synonym for the *normal reaction force* when marking. On the other hand, Marker 4 had been more generous when accepting synonyms for the *normal reaction force*, and even acknowledged in their own marking instances where they had given the benefit of the doubt.

From the above considerations, two questions appeared to be problematic for the human markers; these were Q3 and Q19. In both of these cases, the marking agreement and the Cohen's kappa value between the human markers and the UHM was not good across several cases. The issues with Q3 were found to arise from almost all of the answers being marked as correct, whereas those with Q19 were found to be the result of unclear marking guidance. It is of note that the issues found with marking Q3 could (and did) arise for human and computer markers; whereas the issues raised with the marking of Q19 could only arise with human markers, since the computer marking scheme has no concept of subjectivity.

The UHM was built using 5 markers, with the majority view being taken as the awarded mark for each individual response. As a result, there existed some borderline cases where 3 markers chose to mark the response one way, with the other 2 markers choosing to mark the same response the other way. For triangulation with the above findings, the number of these borderline cases encountered in the human marking of each free-response question on Version 1 of the AMS is given in Table 6.7 below. The questions with a high number of borderline cases were Q19, Q30, and Q32; this is in agreement with the above human marking IRR calculations, since the same questions were identified to have high amounts of disagreement between various human markers and the UHM when marked. In particular, Q19 had a very high number of borderline cases, which is further reflective of the problems encountered by the human markers when interpreting and applying the provided marking scheme on this question. As a result, the marking guidance corresponding to Q19 was modified to give examples of what does and does not count as an acceptable synonym for the *tension* force. In addition, the marking guidance for Q30 was similarly modified to give examples of typical *correct* and *incorrect* answers. In the case of Q32, there were found to be no issues with the marking guidance, and the high number of borderline cases resulted from one marker being overly generous with their marking, as discussed above.

The UHM was designed to be an ultimate marking tool, by employing a *majority rules* approach to harness human marker expertise while reducing the subjective aspect of human marking. However, each of the humans used to build the UHM had to interpret the provided marking guidance in order to mark the responses. For the majority of the questions, each of the human markers interpreted the marking guidance and marked the responses in the intended way, meaning that the subjective aspect of human marking was not a concern in these cases. For a small number of other questions, the interpretation of the marking guidance was found to vary widely from person to person, making the process of developing a UHM highly subjective for these cases. From these considerations, the effectiveness of the UHM as an objective marking construct comes down to whether the different human markers interpret the provided marking guidance in the intended way, which is an unavoidable aspect of human marking which has previously been discussed by Butcher and Jordan (2010). A further discussion of issues pertaining to human markers is given in **Subsection 7.4.6**.

Version 1 AMS free-response question	Number of responses	Number of borderline human marking cases
Q1	328	0
Q2	307	10
Q3	305	10
Q4	304	12
Q5	301	1
Q11	280	4
Q13	277	10
Q17	276	3
Q19	275	76
Q20	275	3
Q22	275	10
Q23	275	4
Q25	255	2
Q27	255	1
Q29	255	3
Q30	254	16
Q31	254	1
Q32	254	21

Table 6.7: Table showing the number of borderline cases encountered in the human marking of each of the free-response questions on Version 1 of the AMS.

6.5 Conclusions

The first aim of the study presented in this chapter was to test the Version 1 AMS questions and marking rules for reliability. To this end, Version 1 AMS response data were collected from students during the academic year 2017-2018. To test the AMS questions for reliability, *Classical Test Theory* (CTT) statistics were calculated on the Version 1 response data set, and it was found that most of the questions on Version 1 of the AMS were functioning well. To test the AMS marking rules for reliability, *Inter-Rater Reliability* (IRR) statistics were calculated on the Version 1 response data set. The Version 1 IRR study found that the computer marking performed well for just over a third of the free-response questions on Version 1 of the AMS, and findings from this study were used to modify the questions and marking rules for Version 2 of the AMS (the next iteration of the AMS in its development process) where required. As a result, the IRR strand of the study indicated that further development and testing of the computer marking was still needed to develop the AMS into a concept inventory suitable for general and widespread use.

The secondary aim of the IRR study was to investigate the facets of using automated marking on the free-response questions in the AMS. It was found that there were advantages and disadvantages in the approach of using false positive and false negative cases to develop the computer marking rules, and limitations of the approach were encountered in questions where it was ineffective; in particular, question wording was highlighted as an important consideration when attempting to develop effective computer marking. In addition, issues concerning the subjectivity of human marking were raised by some questions when testing the UHM for consistency, with different possible interpretations of the marking guidance being identified as a key factor contributing to this. As a result, these findings provided useful general points to consider in the development of automated marking schemes for free-response concept inventory questions.

The findings from the CTT and IRR studies conducted in this chapter were used to iterate Version 1 of the AMS into Version 2, which completed an iterative step in the development process of the AMS. In order to check its level of functionality, Version 2 of the AMS needed to be tested for reliability; this testing is the focus of the next chapter.

6.6 Summary and looking ahead

Chapter 6 presented the quantitative findings from the *Classical Test Theory* (CTT) and *Inter-Rater Reliability* (IRR) studies conducted using responses gathered to Version 1 of the AMS. The findings indicated that the AMS questions were functioning well, but the marking rules still required further development.

Chapter 7 focuses on further testing of the AMS questions and marking rules using quantitative approaches. It presents findings from data gathered through administration of Version 2 of the AMS in the academic year 2018-2019.

7 Expanding the free-response aspect of the Alternative Mechanics Survey

7.1 Rationale

In **Chapter 6**, the Version 1 iterations of the AMS questions and marking rules were tested for reliability using quantitative approaches. The *Classical Test Theory* (CTT) approach was used to test the AMS questions, and it was found that the questions were generally functioning well. In contrast, the *Inter-Rater Reliability* (IRR) approach used to test the AMS marking rules found that the marking rules still required further development and testing in order to function consistently. Alongside this functionality, one of the main aims of the overall research is to develop concept inventories which make use of free-response questions in their construction. In accordance with this aim, seven selected-response questions from Version 1 of the AMS were converted into free-response format for Version 2 of the AMS, thus expanding the free-response scope of the AMS. In order to do this, marking rules from related free-response questions on Version 1 of the AMS were transferred to the new Version 2 free-response questions.

After the Version 1 study, the AMS questions and marking rules were changed from Version 1 to Version 2, with the intention of improving their performance. Since reliability cannot be assumed to carry over after changes have been made, both the questions and marking rules needed to be tested again for reliability in the Version 2 study. As a result, the primary aim of the work presented in the current chapter was to test the reliability of the Version 2 variants of the AMS questions and marking rules by applying the CTT and IRR approaches to the data set collected using the Version 2 of the AMS in the academic year 2018-2019. In addition to this, a secondary aim of the Version 2 IRR study was to test how effective the rule transfer approach was for developing marking rules for the new Version 2 free-response questions.

7.2 Methods

7.2.1 Data collection

The Version 2 AMS questions were put into a Moodle test hosted on the OpenScience Laboratory (OSL), and this was used to collect the Version 2 AMS data set. The relevant approvals were gained from The Open University's *Human Research Ethics Committee* and *Student Research Project Panel*, which allowed data to be collected

from participants; this was done by contacting possible participants with information about the project, along with a link to Version 2 of the AMS on the OSL. Unlike in the Version 1 AMS study, the potential participants were all OU undergraduate students, on this occasion on the OU modules *S112 Science: Concepts and Practice*, *SM123 Physics and Space*, *S383 The Relativistic Universe*, *SM358 The Quantum World*, and *SMT359 Electromagnetism*, with no data gathered from participants from other universities or secondary schools. Once all of the participants had responded to the AMS, the data were downloaded from the OSL. Blank entries were removed, and complete tests were retained for calculation of the CTT statistics; all non-blank entries for each question were separately retained for calculation of the IRR statistics.

Human marking of the responses was again needed to compare against the computer marker, which required a *Unified Human Marker* (UHM) to be constructed for the Version 2 AMS data set. For consistency, the same five markers and approach employed to build the Version 1 UHM (see **Subsection 6.2.1**) were engaged to build the Version 2 UHM; although the marking guidance used by the markers had been updated based on based on problems identified in the Version 1 IRR study.

The new free-response questions on Version 2 of the AMS

In line with the overall aims of the research to develop concept inventories which make use of free-response questions, every question on Version 2 of the AMS was used in the free-response format. This required the selected-response questions from Version 1 of the AMS to be changed into free-response questions for Version 2. In the cases of Q8, Q9, Q10, Q14, Q16, Q24 and Q26, the questions asked test-takers to identify a trajectory, so these questions were converted into FRQ(L) questions, requiring the entry of a single letter corresponding to the multiple-choice option. In the cases of Q7, Q12, Q15, Q18, Q21, Q28 and Q33, the questions were converted into full free-response questions which required the entry of a short phrase or sentence to be answered. As a result, these questions required marking rules, and these were inherited from free-response questions from Version 1 of the AMS which tested similar content and concepts. The question pairs used in this rule transfer are shown in Table 7.1 below.

New Version 2 AMS free-response question	Situation	Version 1 AMS question where rules were inherited from	Situation	Concept Tested
Q7	Marble in track	Q13	Hockey puck	Types of forces
Q12	Hockey puck	Q27	Rocket	Newton's First Law
Q15	Ball toss	Q3	Ball drop	Kinematics (Projectile motion)
Q18	Truck and car	Q17	Truck and car	Newton's Third Law
Q21	Boy on swing	Q19	Elevator	Types of forces
Q28	Woman pushing box	Q20	Elevator	Newton's First Law
Q33	Tennis ball	Q3	Ball drop	Types of forces; Kinematics (Projectile motion)

Table 7.1: Table showing which Version 1 free-response AMS questions the marking rules for new Version 2 free-response AMS questions were inherited from.

In each of the question pairs, the situations being tested were not exactly the same. However, the free-response answers given to each of the Version 1 questions in the pair appeared to be sufficiently similar to the types of answers that would be expected to be given to the new Version 2 question in the pair, which allowed the marking rules to be transferred from one question to the other. For example, Q3 of Version 1 of the AMS asks test-takers to identify the forces acting on a ball after it is dropped, and a correct answer to this question would be *"weight"*. In comparison, Q33 of Version 2 of the AMS asks test-takers to identify the forces acting on a tennis ball after it has been hit, and the correct answer of *"weight"* for Q3 of Version 1 of the AMS would also be a correct answer for Q33 of Version 2 of the AMS, which appeared to make the transfer of marking rules feasible in this case.

Q6 of Version 1 of the AMS was a multiple-response question that asked test-takers to identify the forces acting on a marble while it was inside a frictionless shaft. To answer correctly, test-takers needed to identify three different forces, and no other question on the AMS has this requirement. As a result, there was no free-response question on Version 1 of the AMS from which Q6 could inherit related marking rules, meaning that it could not be converted to free-response format for Version 2 of the AMS. It was therefore decided to exclude Q6 from Version 2 of the AMS, since if it had been included, it would have been the only selected-response question amongst the free-response questions.

7.2.2 Data analysis

The CTT statistics of difficulty, discrimination, point biserial coefficient, Kuder-Richardson Reliability and Ferguson's delta (as detailed in **Subsection 6.2.2**) were calculated for the Version 2 data set in order to evaluate the functionality of the Version 2 AMS questions. The results of these calculations can be found in **Section 7.3**.

The IRR statistics of marking agreement and Cohen's kappa (as outlined in **Subsection 6.2.2**) were calculated using the Version 2 AMS response data in order to check the reliability of the corresponding Version 2 marking rules. The results of the IRR calculations can be found in **Section 7.4**. As for the Version 1 data, the false positive and false negative cases were used to improve the marking rules for each question and the revised rules were back-tested against the responses used to develop them. Further, the individual human markers were tested for consistency against the UHM,

leading to further consideration of the reasons for inaccuracies in human marking and the relative strengths and weaknesses of human and computer marking; this discussion can be found in **Subsection 7.4.6**. The Version 2 AMS questions used to conduct these studies can be found in **Appendix D**.

7.3 Results and Discussion: AMS Version 2 CTT study

7.3.1 Total score and number of attempts

There were 81 respondents to Version 2 of the AMS (Of these, 69 were students on the module *S112*, while the other 12 students were from other modules), and 60 submitted tests that were *complete*, meaning that they had answered all 32 of the items (Of these, 52 were students on the module *S112*, while the other 8 students were from other modules); note that there were 32 items on Version 1 of the AMS as opposed to the 33 items on Version 1 of the AMS. The graph showing the frequency of each of the different scores for the $N = 60$ completed tests is given in Figure 7.1 below, and these are the scores as were awarded by the Version 2 UHM.

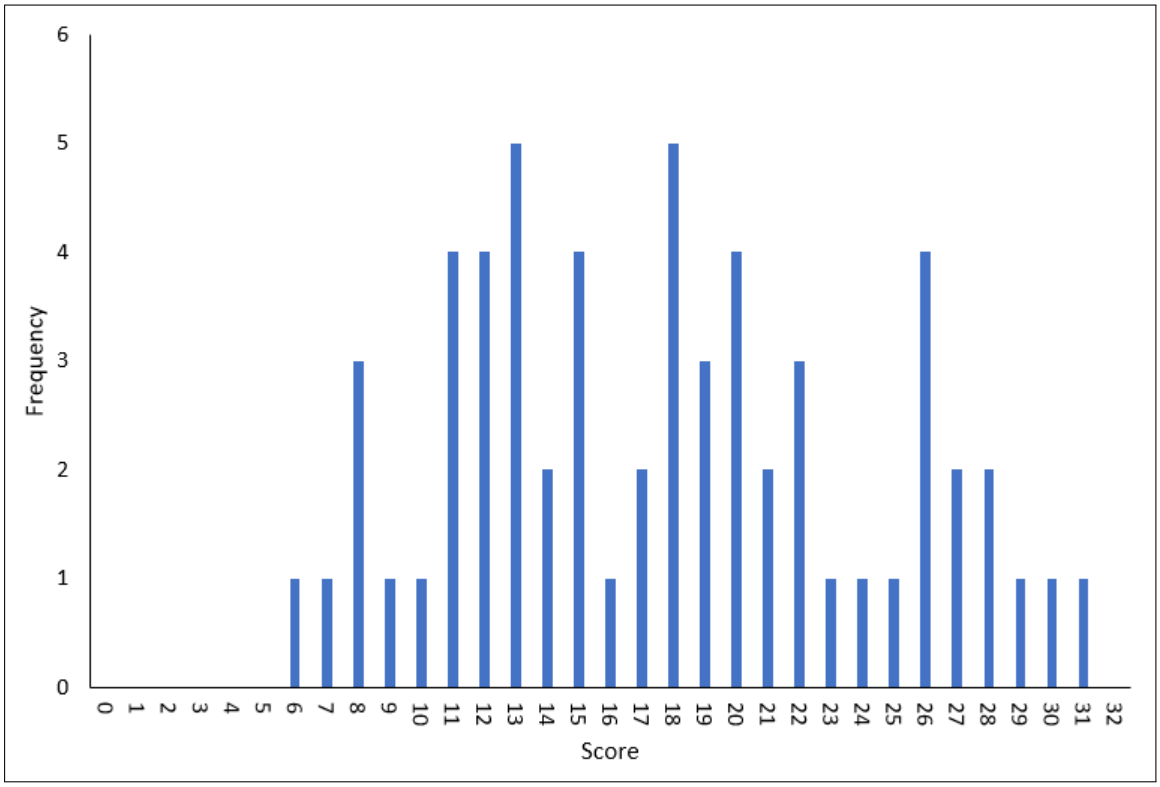


Figure 7.1: Graph showing the distribution of the scores on Version 2 of the AMS for all 60 completed tests marked by the UHM.

It can be observed from Figure 7.1 that the highest score attained on Version 2 of the AMS was 31 out of 32, and the lowest score attained was 6 out of 32. There were two modal scores: 13 and 18. The overall graph did have standard normal distribution features, with few participants attaining the higher and lower scores, and the majority attaining scores toward the middle of the distribution. The mean score on the test was 17.70, which is slightly above the middle value of 16; this is of significance because if each item on the AMS had the optimal difficulty of 0.5, then the mean score would be expected to be the middle value of 16.

The respondents were made up of Open University distance learning undergraduate students only, with the majority of the students being taken from the introductory science course *S112 Science: Concepts and Practice*. Since Open University undergraduate modules are open entry, and this is an OU level one module, the respondents will have had a variety of previous exposure to Newtonian mechanics. Also, the way in which modules combine into qualifications means that very few students on *S112* will be intending to take a physics or astronomy pathway, further reducing the likelihood of familiarity and interest with basic physics concepts. These factors may explain the variability in the scores shown in Figure 7.1 and also explain why the mean score for the Version 2 data is lower than the mean score for Version 1 (as presented in **Subsection 6.3.1**).

Test-takers had the option of abandoning the AMS at any time, and 21 of the respondents did not complete the AMS. Figure 7.2 shows how far through the test each of the 81 participants managed to get before giving up, and $N = 60$ respondents completed the entire AMS. Note that Q6 of Version 1 of the AMS was not present in Version 2, but the *standardized AMS question numbering* (the same as the Version 1 numbering, see Tables A.1 and A.2 in **Appendix A**) is used throughout this thesis for ease of comparison between different versions of the AMS.

Figure 7.2 shows the number of attempts that were made on each question on Version 2 of the AMS. There was a sharp decrease from Q1 to Q2, possibly because some respondents only looked at Q1 before deciding not to engage further with the rest of the AMS. There was then a steady decrease down to Q9, before another sharp decrease to Q10, where the graph leveled out until Q16. There was further decrease to Q17, where the graph again leveled out until Q21. At Q22, there was one last decrease, and there were no further decreases thereafter. The sharp decrease between Q9 and

Q10 was not observed in Version 1 of the AMS. Q10 is the first in a sequence of four questions about the motion of a hockey puck, so respondents may have chosen to give up here because of the amount of reading and effort required to answer the four-part question. On the other hand, the decrease between Q21 and Q22 was also observed in Version 1 of the AMS, although it occurred at Q23 in Version 1 instead of at Q22. As previously discussed in the context of the Version 1 AMS findings in **Subsection 6.3.1**, it is possible that students with a lower previous exposure to physics found the *moving blocks* scenario to be difficult, leading them to abandon their AMS attempts. This idea can be applied to explain the drop-off between Q21 and Q22 in the Version 2 findings. Since *S112* students typically do not have a physics background and do not want to study physics, they may be expected to abandon their AMS attempt when confronted with the *moving blocks* scenario.

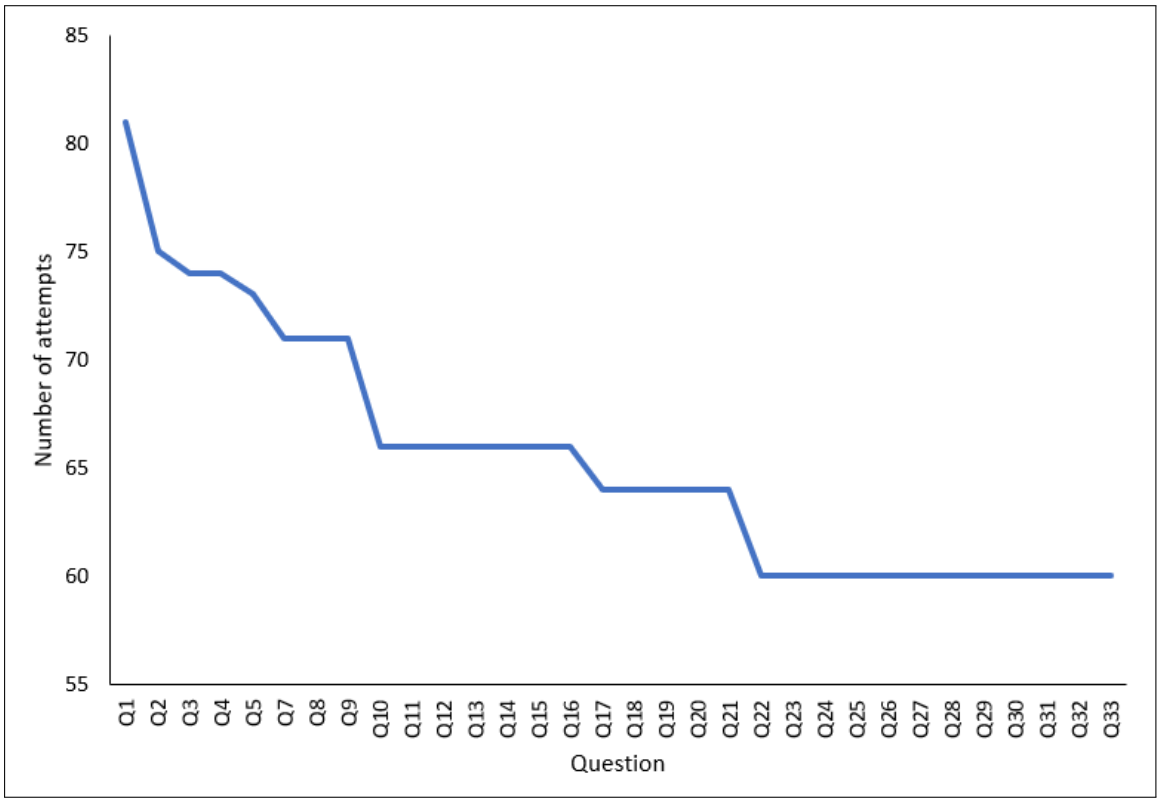


Figure 7.2: Graph showing the number of attempts made on each Version 2 AMS question.

Comparison of the Version 1 and Version 2 cohorts

The populations from which the samples were drawn were different for the Version 1 and Version 2 cohorts. The Version 1 sample was drawn from a combination of high school physics students, OU undergraduate students from various modules, and undergraduate students from an external university; the majority of these students had encountered Newtonian mechanics in some detail before taking Version 1 of the AMS. By contrast, the Version 2 sample was drawn mostly from OU undergraduate students from the level 1 module *S112 Science: Concepts and practice*; since these students are not likely to be registered on physics or astronomy pathways, and have not studied Newtonian mechanics in their OU modules, it follows that they are less likely to have had previous exposure to Newtonian mechanics. Consistent with these observations, the mean score on Version 1 of the AMS was higher than that on Version 2 of the AMS. In addition, the data pertaining to the total scores on Version 1 was left-skewed, whereas the corresponding total scores data on Version 2 had standard normal distribution features. In this respect, the Version 1 cohort could be considered as akin to a *post-test* cohort, whilst the Version 2 cohort more resembles a *pre-test* cohort. Furthermore, the difference in experience between the cohorts meant that the AMS questions and marking rules were tested with a variety of test-takers, which is important since the AMS could conceivably be used with a broad range of different student cohorts.

7.3.2 Difficulty and dynamic difficulty

The *difficulty* and *dynamic difficulty* are calculated here, as previously defined in **Subsection 6.2.2**. Data relating to these two types of difficulty are presented in Table 7.2 and Figure 7.3 below. Note that the data used for these calculations were the Version 2 AMS responses marked by the UHM. Further note that every question on Version 2 of the AMS was in free-response format. As a result, FRQ is used to denote a free-response question that has not changed question type from the previous AMS version; FRQ(L) is used to denote a free-response question that requires a single letter entry; and FRQ(NEW) is used to denote a free-response question that was a different question type in the previous AMS version.

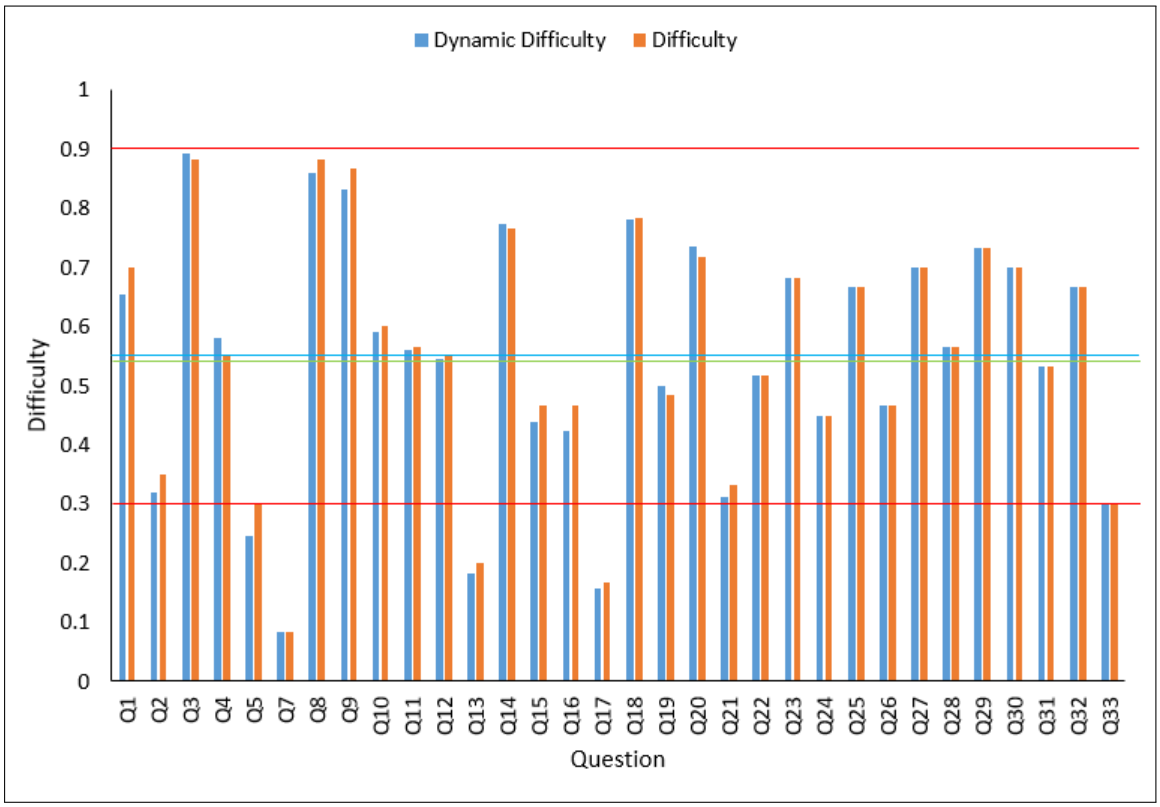


Figure 7.3: Graph showing the dynamic difficulty (blue) and difficulty (orange) of each question on Version 2 of the AMS. The red horizontal lines indicate the lower and upper bounds of the acceptable range of values for the difficulty; the blue horizontal line indicates the mean value of the difficulty; and the green horizontal line indicates the mean value of the dynamic difficulty. Note that higher values indicate *easier* items, whereas lower values indicate *harder* items.

Question	Question type	Dynamic Difficulty	Difficulty
Q1	FRQ	0.65	0.70
Q2	FRQ	0.32	0.35
Q3	FRQ	0.89	0.88
Q4	FRQ	0.58	0.55
Q5	FRQ	0.25*	0.30
Q7	FRQ(NEW)	0.08*	0.08*
Q8	FRQ(L)	0.86	0.88
Q9	FRQ(L)	0.83	0.87
Q10	FRQ(L)	0.59	0.60
Q11	FRQ	0.56	0.57
Q12	FRQ(NEW)	0.54	0.55
Q13	FRQ	0.18*	0.20*
Q14	FRQ(L)	0.77	0.76
Q15	FRQ(NEW)	0.44	0.47
Q16	FRQ(L)	0.42	0.47
Q17	FRQ	0.16*	0.17*
Q18	FRQ(NEW)	0.78	0.79
Q19	FRQ	0.50	0.48
Q20	FRQ	0.73	0.72
Q21	FRQ(NEW)	0.31	0.33
Q22	FRQ	0.52	0.52
Q23	FRQ	0.68	0.68
Q24	FRQ(L)	0.45	0.45
Q25	FRQ	0.67	0.67
Q26	FRQ(L)	0.47	0.47
Q27	FRQ	0.70	0.70
Q28	FRQ(NEW)	0.57	0.57
Q29	FRQ	0.73	0.73
Q30	FRQ	0.70	0.70
Q31	FRQ	0.53	0.53
Q32	FRQ	0.67	0.67
Q33	FRQ(NEW)	0.30	0.30

Table 7.2: Table showing the dynamic difficulty and difficulty of each question on Version 2 of the AMS.

The dynamic difficulty was larger than the difficulty for Q3, Q4, Q14, Q19 and Q20. This means that the total number of respondents who attempted these questions found them easier on the whole than the number of respondents who answered all of the questions. The dynamic difficulty was equal to the difficulty Q7, meaning that this question was of the same difficulty for test-takers who submitted partially complete attempts, and for test-takers who submitted complete attempts. Further, the dynamic difficulty was equal to the difficulty from Q22 onwards, because in these cases the total number respondents who attempted the questions was equal to the total number

of respondents who answered all of the questions. For the rest of the questions, the dynamic difficulty was smaller than the difficulty, meaning that the total number of respondents who attempted these particular questions found them harder on the whole than the number of respondents who attempted all of the questions. As mentioned previously in **Subsection 6.3.2**, the dynamic difficulty value is generally expected to be less than or equal to the corresponding difficulty value for a question. This trend was observed in the majority of the questions, but there were some cases where the opposite trend was observed, and these are discussed below.

Cases where the dynamic difficulty was greater than the difficulty

The number of test-takers who answered Q19 and Q20 was 64, and the number of test-takers who answered Q14 was 66; in contrast, the number of test-takers who answered all of the questions was 60. Hence the number of test-takers who answered these questions was slightly different from the number of test-takers who answered all of this questions, which means that dynamic difficulty and difficulty values of these questions would be expected to be slightly different, as a result of stochastic effects, as in **Subsection 6.3.2**. In the cases Q3 and Q4, the same stochastic process can also be used to explain the effect. This is because the effect of the dynamic difficulty being higher than the difficulty is small in these questions, which means that the effect is within what might be reasonably expected from random fluctuations in the data.

In the cases where the dynamic difficulty was larger than the difficulty, the difference was never greater than 0.03. This effect was larger than in the cases identified from Version 1 of the AMS, but there were fewer participants in the Version 2 study, which amplifies the effect. As was the case in the Version 1 study, it is also useful to look at cases where the difficulty is outside the acceptable range of values, or where it is close to the boundaries of this acceptable range. As mentioned previously, the acceptable range of values for the difficulty are [0.3, 0.9]. For the upper bound, three questions on the AMS had a difficulty values that were close to the cut-off of 0.9; for the lower bound, four questions had difficulties values that were equal to or below the 0.3 boundary. Note that there were more difficulty-based problem cases on Version 2 of the AMS than there were for Version 1.

Cases where the difficulty values were high

Q3 had difficulty and dynamic difficulty values that were close to the 0.9 boundary, meaning that almost all of the test-takers who attempted the question got it right. The question asks test-takers to identify the forces acting on a stone after it has been dropped from a building. This question corresponds to Q3 of Version 1 of the AMS, and the pattern observed with respect to difficulty values is the same. It was retained from the previous version because of its link to Q4, although it may now require further modification or removal since it has twice been flagged as a problematic item when difficulty is used as the criterion.

Q8 had difficulty and dynamic difficulty values that were slightly below the cut-off of 0.9. Q8 requires test-takers to select the trajectory of a marble after it has been shot out of a smooth track. It corresponds to Q8 of Version 1 of the AMS, and the pattern observed in the difficulty values is the same in both versions. It is of note that Q8 question is posed with a free-response answer box instead of multiple-choice options in Version 2, although this does not appear to have affected the difficulty of the question.

Q9 was the third question with higher values for difficulty and dynamic difficulty, illustrating that it was an easier question for the Version 2 cohort. It corresponds to Q9 of Version 1 of the AMS, and the pattern observed is different in this case, since the difficulty and dynamic difficulty values for the Version 1 variant of the question were not close to the upper bound of the acceptable range of values. It is a question that asks test-takers to identify the trajectory of a steel ball after being thrown as a hammer. The question gives the trajectories to choose from, and has corresponding multiple-choice options in the Version 1 variant, whereas the Version 2 variant requires students to enter the trajectory as a letter into a free-response box. Unlike in the case of Q8, changing the question format appears to have made Q9 easier. However, differences between the Version 1 and Version 2 cohorts may be a more likely explanation for this effect.

Cases where the difficulty values were low

Q7 was the hardest question on the AMS in terms of difficulty and dynamic difficulty. It was found to be particularly difficult for the Version 2 cohort, with only 6 correct answers being given out of 72 attempts at the question. The question asks

students to identify the forces acting on a marble after it emerges from a frictionless channel, and it was changed from being a multiple-response question on Version 1 of the AMS to a free-response question on Version 2 of the AMS. It is possible that changing the format of this question contributed to the increase in difficulty, since test-takers now have to identify the forces *ab initio* for themselves, rather than select them from a pre-prepared list.

Examining the responses revealed two main classes of incorrect answer. In the first class, students referred to a *centripetal force* acting; in the second, students referred to a *thrust* force acting. Misconceptions about the *centripetal force* have been identified in the literature (Yasuda et al., 2018; Rebello and Zollman, 2004), and misconceptions about the presence of active forces are also common (Eaton et al., 2019); taking these together with the potential inexperience of the cohort provides a possible explanation as to why Q7 was particularly difficult. This case illustrates the importance of considering students' incorrect answers as well as their correct answers, a point which has previously been highlighted in the literature by Dedic et al. (2010) and Smith et al. (2020).

Q17 had difficulty and dynamic difficulty values that were below the 0.3 boundary, and it was the second-hardest question on the AMS in terms of these statistics. The question involves the situation of a truck pushing a car, and requires test-takers to compare the forces acting on the truck and car during this motion. It corresponds to Q17 of Version 1 of the AMS, which was also the third hardest question in that version of the AMS, and the question wording is the same as in the previous version. However, Q17 tests the concept of Newton's Third Law, and the above findings add further credence to the idea that this concept is difficult for students in general to understand and master. This was further reflected in the incorrect responses given to the question, which often referred to one of the two forces involved (either *the force of the truck on the car* or *the force of the car on the truck*) being greater than the other force.

The difficulty and dynamic difficulty values of Q13 were both below the acceptable value of 0.3. The question asks test-takers to identify the forces acting on a hockey puck as it travels along a frictionless surface. There were three different types of incorrect answer given to this question: those which mentioned weight, but missed the normal reaction force; those which added an extra applied force; and those which

identified energy as a force. As a result, Q13 was answered poorly by the majority of the cohort, an effect which was not observed for the corresponding Q13 on Version 1 of AMS. This is consistent with the findings of Martin-Blas et al. (2010), who found that that students with a lower previous exposure to physics (such as the Version 2 cohort) were less than half as likely to get FCI Q11 (which corresponds to AMS Q13) right as those with higher previous exposure to physics (such as the Version 1 cohort).

Q5 had a dynamic difficulty value that was below the 0.3 boundary, whereas its difficulty value was on this boundary. It corresponds to Q5 of Version 1 of the AMS, and it was not found to be particularly hard for test-takers in the Version 1 cohort. Q5 asks test-takers to compare the forces acting on a truck and a car during a collision. It tests the concept of Newton's Third Law, which is a possible explanation as to why the Version 2 cohort struggled with the question. Alternatively, since this effect was not observed in the Version 1 AMS data, it is also possible that this question is difficult specifically for the Version 2 cohort. This possibility was also reflected in the Version 2 cohort's responses to this question. There was not any characteristic or fundamental differences in the structure or content of the responses given by the Version 1 and Version 2 cohorts to this question, but the Version 2 cohort simply gave more incorrect responses.

The effect of changing question type on difficulty

Seven questions that were not free-response format on Version 1 of the AMS were converted to full free-response versions (not FRQ(L) questions) on Version 2 of the AMS, and taking this step appeared to affect the difficulty for most of these questions. For Q18 and Q28, there was little change in the difficulty value between the two versions; whereas for Q7, Q12, Q15, Q21 and Q33, the free-response versions of the questions were harder than the versions that were not free-response. This finding may indicate that free-response versions of questions are typically more difficult than their multiple-choice or multiple-response counterparts. However, aforementioned differences between the previous exposure to Newtonian mechanics of the Version 1 and Version 2 cohorts are likely to have influenced the difficulties of the questions; this means that based on these findings alone, the free-response versions of the questions cannot be assumed to be harder than versions that are not free-response in general.

Summary

Overall, 28 out of the 32 questions had a difficulty value and a dynamic difficulty value that were within the acceptable range of [0.3, 0.9]. The mean value of the difficulties of the individual questions was 0.55, and the mean value of the dynamic difficulties of the individual questions was 0.55. Both of these mean values were within the acceptable range of values for difficulty, which implied that the AMS questions functioned in the desired way overall when difficulty was used as the metric. In line with the objectives of the over-arching study, the questions flagged by the above difficulty considerations were considered and revised as appropriate for the next version of the AMS.

7.3.3 Discrimination and point biserial coefficient

The discrimination and point biserial coefficient values were calculated using the Version 2 AMS responses marked by the UHM. The results are shown in Table 7.3 and Figure 7.4 below.

Question	Question type	Discrimination	Point biserial coefficient
Q1	FRQ	0.53	0.37
Q2	FRQ	0.73	0.61
Q3	FRQ	0.27*	0.28
Q4	FRQ	0.47	0.33
Q5	FRQ	0.60	0.47
Q7	FRQ(NEW)	0.27*	0.31
Q8	FRQ(L)	0.47	0.35
Q9	FRQ(L)	0.40	0.25
Q10	FRQ(L)	0.80	0.61
Q11	FRQ	0.73	0.47
Q12	MCQ	0.60	0.35
Q13	FRQ	0.67	0.61
Q14	FRQ(L)	0.67	0.41
Q15	FRQ(NEW)	0.53	0.44
Q16	FRQ(L)	0.80	0.55
Q17	FRQ	0.33	0.32
Q18	FRQ(NEW)	0.67	0.43
Q19	FRQ	0.40	0.40
Q20	FRQ	0.73	0.43
Q21	FRQ(NEW)	0.87	0.67
Q22	FRQ	0.53	0.33
Q23	FRQ	0.47	0.36
Q24	FRQ(L)	0.87	0.54
Q25	FRQ	0.87	0.51
Q26	FRQ(L)	0.93	0.61
Q27	FRQ	0.67	0.53
Q28	FRQ(NEW)	0.73	0.38
Q29	FRQ	0.27*	0.17*
Q30	FRQ	0.53	0.31
Q31	FRQ	0.60	0.39
Q32	FRQ	0.80	0.57
Q33	FRQ(NEW)	0.80	0.62

Table 7.3: Table showing the discrimination and point biserial coefficient values for each question on Version 2 of the AMS.

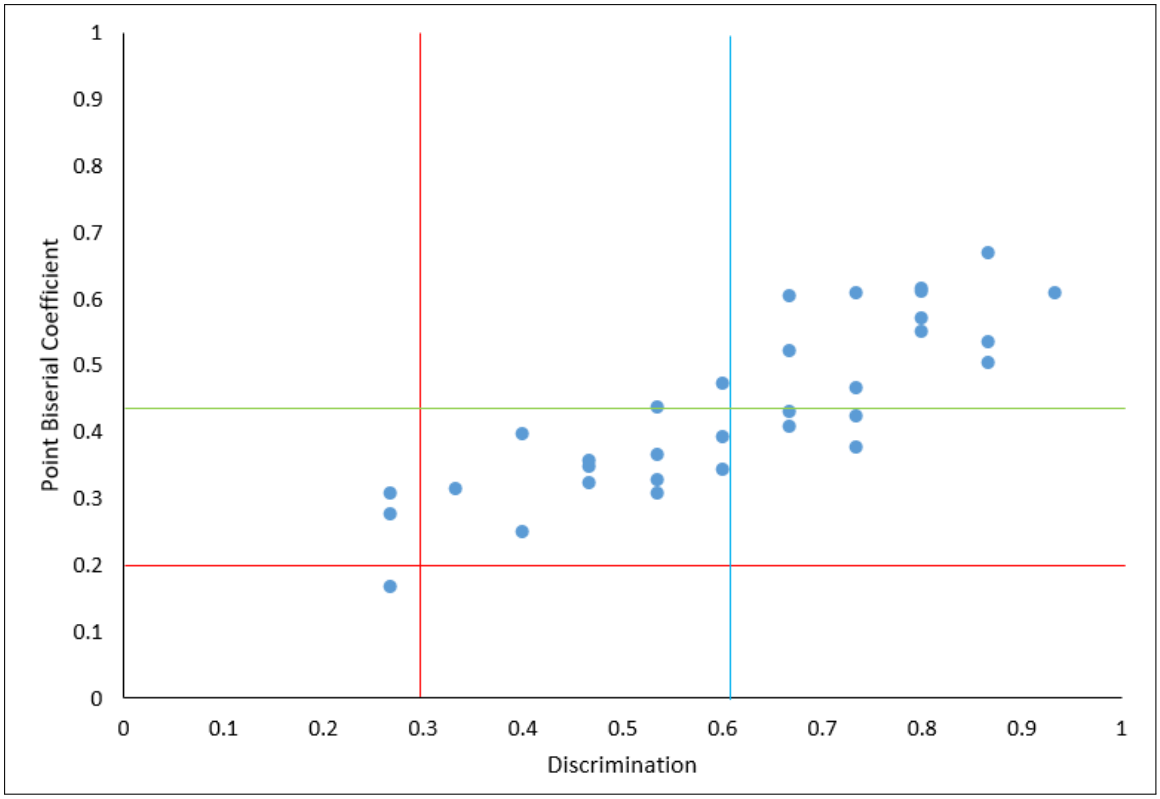


Figure 7.4: Graph showing the point biserial coefficient and the discrimination of each question on Version 2 of the AMS. The red horizontal line represents the lower bound of the acceptable values for point biserial coefficient, and the green horizontal line represents the mean value of the point biserial coefficient. The red vertical line represents the lower bound of the acceptable values for discrimination, and the blue vertical line represents the mean value of the discrimination.

The acceptable range of values for discrimination are [0.3, 1], and the acceptable range of values for point biserial coefficient are [0.2, 1]. One question had discrimination and point biserial coefficient values that were both outside of the acceptable ranges; two questions had discrimination values that were outside the acceptable range; and one question had a particularly high value for the discrimination. These cases are discussed here.

Cases where the discrimination and/or point biserial coefficient were low

Q29 had discrimination and point biserial values that were outside the acceptable range of values. It is a free-response version of Q29 from Version 1 of the AMS, where it was previously a multiple-choice question. The Version 1 variant of this question also had lower values for the discrimination and point biserial coefficient statistics, so changing the format of the question does not seem to have affected the values of these statistics in this case. The question itself asks test-takers to describe what happens to

the speed of a moving box when the force applied to it is doubled, with the correct answer being that the speed of the box *increases*.

From the responses, it was found that most of the incorrect answers referred to the speed of the box *doubling*, with the other incorrect answers referring to the speed increasing *exponentially* or *at a constant rate*. The responses indicate that most of the test-takers had some understanding of the situation, but those who answered incorrectly were unable to identify that force and speed are not related in the same way that force and acceleration are. Taking this together, the low discrimination can be explained because the question was differentiating between a population with full understanding of the concept, and a population with partial understanding, with there being little to no population with limited or no understanding; this was hence an effect related to the characteristics of the Version 2 cohort. This lack of discriminatory power additionally provides a possible explanation for the low point biserial coefficient value, as the population tested had a higher level of understanding of Q29 than would be expected, whereas the population typically demonstrated a more varied level of understanding in the majority of the other questions on the AMS.

Q3 had a discrimination value that was outside the acceptable range of values, whereas its point biserial coefficient value was at the lower end of the acceptable range of values. Q3 asks test-takers to identify the forces acting on a stone after it has been dropped from a building. It corresponds to Q3 of Version 1 of the AMS, and the Version 1 variant of the question had discrimination and point biserial coefficient values that were outside the acceptable range of values, meaning that Q3 was problematic in terms of these statistics in both Version and Version 2 of the AMS.

The discrimination value of Q7 was outside the acceptable range of values, whereas its point biserial coefficient value was acceptable. The question asks test-takers to identify the forces acting on a marble after it has been shot out of a frictionless channel. It corresponds to Q7 of Version 1 of the AMS, and the Version 1 variant of the question did not have issues with its discrimination and point biserial coefficient values. The question wording is the same in both versions, but the question format was changed from multiple-response in Version 1 to free-response in Version 2, which could explain the lower values in Version 2. However, differences between the Version 1 and Version 2 cohorts are also likely to have contributed to this effect. Q7 was previously identified to have difficulty issues in **Subsection 7.3.2**, since almost everybody who attempted

the question got it wrong. This is the reverse of the effect observed for Q3, where everybody who attempted the question got it right, but the net effect here is the same: the item is unable to differentiate between higher-scoring and lower-scoring students by definition, and it has a lower discrimination value as a result.

Cases where the discrimination and/or point biserial coefficient were high

Q26 was the question on the AMS with the highest discrimination value. It requires test-takers to select the path taken by a rocket while it is floating in outer space. Q26 of Version 2 of the AMS corresponds to Q26 of Version 1 of the AMS, and the previous version did not have an especially high value for the discrimination coefficient. The question wording is the same in both versions, but the question format was changed from multiple-choice in Version 1 to free-response in Version 2. However, Q26 was changed from a multiple-choice question to a FRQ(L) question, meaning that it was still essentially a multiple-choice question. This was reflected in the responses, which were typically only single-letter entries. It follows that the high discrimination value held by this question could be a result of the change in question format, or it could be caused by a cohort effect. The latter of these is the more likely explanation, since the change to the question format was small (multiple-choice to FRQ(L) rather than multiple-choice to outright FRQ).

Summary

From the above calculations of the difficulty, discrimination and point biserial coefficient statistics, possible issues were raised by AMS items with CTT statistics that were outside the acceptable range of values. In these cases, the issues were often attributed to three main factors. First, the possibility of a stochastic effect caused by random fluctuations in the Version 2 data set, emphasized by the smaller sample size, provided an explanation for the uncharacteristically low or high values of some of the CTT statistics. Second, the question type was changed from selected-response to free-response for some of the questions on Version 2 of the AMS, and this appeared to affect the difficulty of some of these items; this could also have led to changes in the values of the other CTT statistics. Third, a cohort effect was identified as a possible cause of the differences between the Version 1 and Version 2 CTT statistics. This is because the Version 2 cohort were mainly (52 out of 60 of the completed attempts) *S112* students, meaning that they would be expected to have less understanding of

Newtonian mechanics than the Version 1 cohort (which contained mainly students studying physics), as they are intending to do other sciences and only taking some introductory physics. However, the latter of these is likely to be the dominating factor here, since the difference between cohorts is an effect which is consistent throughout all of the AMS questions.

Overall, 29 out of the 32 questions had discrimination values that were within the acceptable range of values. This meant that on the question level, 29 out of the 32 questions could differentiate between higher and lower performing students. For the test level statistics, the mean of the discrimination values of the individual questions was 0.61, which was within the acceptable range for the discrimination value. Taking this together with the question-level findings, this implied that the AMS questions considered as a whole were capable of distinguishing between the higher-performing and lower-performing students. Furthermore, all 32 questions had point biserial coefficient values that were within the acceptable range of values. This meant that on the question level, all of the questions tested concepts that were related to one another. For the test level statistics, the mean value of the point biserial coefficients of the individual questions was 0.44, and this was within the acceptable range of values for the point biserial coefficient. Taking this together with the question-level findings, this implied that the AMS overall contained questions that assessed similar topics. These discrimination and point biserial results provided important evidence for the overall functionality of the AMS questions.

7.3.4 Overall functioning

The mean values for the difficulty, discrimination and point biserial coefficient over all of the questions were calculated, along with the Kuder-Richardson reliability and Ferguson's delta values for the entire test. The results are shown in Table 7.4 below.

A Kuder-Richardson reliability value of 0.87 was calculated for Version 2 of the AMS by using the previously calculated difficulty values for each of the items, the standard deviation of the total scores, and $K = 32$ for the 32 AMS items. This was above the threshold value for Kuder-Richardson reliability of 0.7, which showed that the AMS was reliable overall. A Ferguson's delta value of 0.98 was found for the AMS, by taking the values of the frequency for each of the possible scores, $N = 60$ for the total number of test-takers, and $K = 32$ for the number of test items. This was above

the threshold value for Ferguson's δ of 0.9, and this showed that the overall AMS could be used to discriminate between lower-scoring and higher-scoring students. In addition, the results of this δ calculation agreed with what was concluded from the analysis of the discrimination values from each of the questions previously.

Classical Test Theory Statistic	Value	Desired values
Mean difficulty	0.55	[0.3, 0.9]
Mean discrimination	0.61	≥ 0.3
Point biserial coefficient	0.44	≥ 0.2
Kuder-Richardson reliability	0.87	≥ 0.7
Ferguson's delta	0.98	≥ 0.9

Table 7.4: Table showing the overall CTT statistics for Version 2 of the AMS.

The mean difficulty was within the acceptable range of values for difficulty, meaning that Version 2 of the AMS was overall not too easy nor too hard for the test-takers. In addition, the mean discrimination was within the acceptable range of values for discrimination, and the Ferguson's δ value was also within the acceptable range of values; this showed that the AMS questions were able to distinguish between test-takers of higher and lower performance levels. Furthermore, the mean value of the point biserial coefficient was within the acceptable range, which demonstrated that the items on Version 2 of the AMS tested related content. In addition, the Kuder-Richardson reliability was within the acceptable range of values, which showed that the AMS questions were reliable when considered as a whole. Taken together, these results showed that the Version 2 AMS questions were functioning at an acceptable level. Possible issues were raised by the small number of items with statistics that were outside the acceptable range of values, and changes were suggested for these items. This was in line with the objectives of the overall study, as better functioning AMS questions provide a better idea of students' conceptual understanding of the Newtonian mechanics topics being tested.

7.4 Results and Discussion: AMS Version 2 IRR study

7.4.1 Marking agreement and Cohen's kappa

The marking agreement and Cohen's kappa values were calculated for the Version 2 UHM against the Version 2 computer marking rules, with the results given in Table 7.5 below; the table also gives the number of times the UHM disagreed with the computer marker, and the nature of these disagreements. As was the case in Table 6.4 previously, a *false positive* refers to instances where the computer marked the answer as *correct* and the UHM marked the answer as *incorrect*; whereas a *false negative* refers to instances where the UHM marked the answer as *correct* and the computer marked the answer as *incorrect*.

Question	Number of responses	Number of disagree-ments	Number of false positives	Number of false negatives	Marking agree-ment	Cohen's kappa
Q1	81	0	0	0	1.00	1.00
Q2	75	0	0	0	1.00	1.00
Q3	74	7	6	1	0.91*	0.32*
Q4	74	5	4	1	0.93*	0.86
Q5	73	2	2	0	0.97	0.93
Q7(NEW)	71	5	5	0	0.93*	0.67*
Q11	66	10	8	2	0.85*	0.69*
Q12(NEW)	66	7	4	3	0.89*	0.79*
Q13	66	8	8	0	0.88*	0.68*
Q15(NEW)	66	28	27	1	0.58*	0.22*
Q17	64	0	0	0	1.00	1.00
Q18(NEW)	64	12	2	10	0.81*	0.55*
Q19	64	18	14	4	0.72*	0.44*
Q20	64	2	2	0	0.97	0.92
Q21(NEW)	64	11	5	6	0.83*	0.59*
Q22	60	2	2	0	0.97	0.93
Q23	60	5	1	4	0.92*	0.82
Q25	60	3	2	1	0.95	0.89
Q27	60	2	1	1	0.97	0.92
Q28(NEW)	60	7	7	0	0.88*	0.75*
Q29	60	4	2	2	0.93*	0.83
Q30	60	11	10	1	0.82*	0.49*
Q31	60	1	0	1	0.98	0.97
Q32	60	6	4	2	0.90*	0.77*
Q33(NEW)	60	26	26	0	0.55*	0.25*

Table 7.5: Table showing the number of times the UHM disagreed with the Version 2 computer marking on the Version 2 free-response AMS questions and the nature of these disagreements, as well as the corresponding marking agreement and Cohen's kappa values for the UHM against the Version 2 computer marking.

Table 7.5 contains only those questions which were fully free-response, so those which required students to type a word, phrase or sentence, as IRR calculations are only meaningful when the marking is subjective; FRQ(L) questions are excluded because the marking of these questions is objective. Note also that some of the questions were being used in the free-response format for the first time in Version 2 of the AMS, and these questions are denoted as (NEW) in the *Question* column.

The acceptable range of values for marking agreement are [0.95, 1], whereas the acceptable range of values for Cohen's kappa are [0.8, 1]. A quick examination of Table 7.5 shows that many of the questions had marking agreement and Cohen's kappa values that were outside the respective acceptable ranges. However, it is important to recall that some of these questions were being used in the free-response format for the first time, meaning that their marking rules would not be expected to have a high level of functionality without further development. In addition, there was a lower number of responses to Version 2 than Version 1 of the AMS, and having smaller numbers can cause both positive and negative agreement effects to be inflated when calculating IRR statistics. The different scenarios that emerged from the IRR results are discussed below, with observations and points discussed as relevant.

Cases with acceptable values for both marking agreement and Cohen's kappa

In the cases of Q1, Q2, Q5, Q17, Q20, Q22, Q25, Q27 and Q31, the marking agreement and Cohen's kappa values were each within the acceptable range of values, meaning that there were no concerns about the functionality of the marking rules on these questions. Comparing this with the corresponding findings from the Version 1 IRR study (detailed in **Subsection 6.4.1**), the values of both the marking agreement and Cohen's kappa increased from the Version 1 study for questions Q1, Q2, Q17, Q20, Q22 and Q27. In the case of Q5, the marking agreement increased but the Cohen's kappa value was slightly less than in the Version 1 study; this could be a stochastic effect arising from the smaller Version 2 response set. In the cases of Q25 and Q31, the values of both the marking agreement and Cohen's kappa decreased from the Version 1 study. However, the values of the Version 1 IRR statistics were very high for these questions, which meant that improvement in performance would have been difficult. However, when making these comparisons, it is important to note that the Version 1 and Version 2 cohorts were characteristically different. This means that they would be

expected to give different answers to the AMS questions, and would require different marking rules in order to be accurately marked.

Returning to the Version 2 IRR findings, Q1, Q2 and Q17 had no cases of disagreement between the UHM and computer marking, indicating that the marking rules were operating particularly effectively for these questions. For the other questions, instances of false positives and false negatives arose, and these cases were the same types as those encountered in the Version 1 study; these were previously detailed in **Subsection 6.4.1**. In addition, the same approach used to develop the marking rules in the Version 1 study (discussed in detail in **Subsection 6.2.2** and **Subsection 6.4.2**) was used again in the Version 2 study.

Using false positives and false negatives to develop the marking rules was a general strategy used throughout all of the questions. However, there were other general points about the rule development process which are worth highlighting here. For example, the marking rules for Q4 were modified to account for variations of the root word *"great"* (such as *"greater"*), with a similar approach used for Q31 to account for variations of the root word *"react"* (such as *"reaction"*). In addition, it is often not necessary to develop negative marking rules to mark answers as incorrect; this is because the majority of incorrect answers should simply not be covered by the positive marking rules which mark answers as correct. These examples illustrate that there exist a variety of general strategies that can be applied when developing automated marking schemes, a point which has previously been highlighted in the literature by Miesekes and Pado (2019).

Cases with acceptable values for Cohen's kappa, but lower values for marking agreement

For Q4, Q23 and Q29, the Cohen's kappa values were acceptable, but the corresponding making accuracy values were just below the acceptable range of values. Since the Cohen's kappa values for these questions are above the threshold, this indicates that the marking rules are functioning in the expected way, for reasons outlined in **Subsection 6.4.1**. Comparing with the findings from the Version 1 IRR study (found in **Subsection 6.4.1**), the values of both the marking agreement and Cohen's kappa improved for Q4 and Q23, whereas the Cohen's kappa value improved for Q29; this is encouraging outcome for the development of the marking rules on these questions.

There were false positive and false negative cases for each of questions Q4, Q23 and Q29, as might have been expected for questions in development, especially given the data were from a cohort that appears to have been characteristically different from the one tested previously. However, no serious concerns were found when using the false positive and false negative cases to modify the marking rules of these questions. Taken together with the acceptable values for Cohen's kappa explained above, this was taken to imply that there were no serious concerns about the functionality of the marking rules for these questions, although further testing was required to verify that the changes made to the marking rules were effective.

Cases with lower values for marking agreement and Cohen's kappa

Q3, Q11, Q13, Q19, Q30 and Q32 all had Cohen's kappa and marking agreement values that were outside the acceptable range of values, and in most of these cases the Cohen's kappa values are well below the threshold. This implied that there were problems with the performance of the marking rules on these questions, and this is discussed in what follows. Q3 is known to be a problem case, and its issues arise because it is answered correctly by almost all test-takers that attempt it. In the cases of Q30 and Q32, the value of Cohen's kappa increased since the Version 1 study, which is expected behaviour for questions in development. Q30 was previously identified as a problematic case in the Version 1 IRR study, and these issues evidently carried over to the Version 2 study. In the case of Q32, an unfamiliar type of false positive (caused by some students giving their answers over two lines, leading to human markers to miss parts of the answer) contributed to the lower values of the IRR statistics in the Version 2 study. However, further development and testing was clearly needed to improve the effectiveness of the marking rules on these questions.

In the cases of Q11, Q13 and Q19, the values for the marking agreement and Cohen's kappa were considerably below the acceptable values, and these values were lower than they were in the previous Version 1 IRR study. The Version 2 cohort was smaller than the Version 1 cohort, and this was likely to be a factor contributing to this outcome. However, it is unlikely that this was the only factor influencing this. As previously mentioned, the Version 2 cohort was made up mostly of students on the OU module *S112*, implying that they did not intend on studying physics beyond this module. In contrast, the Version 1 cohort contained mostly students who were studying physics at high school, or who were undergraduate students who were likely

to have had previous exposure to physics. This means that the Version 1 and Version 2 cohorts were characteristically different from one another, and this is likely to have been the main factor influencing the lower IRR statistics in the Version 2 study, since the students from these cohorts are likely to have given different characteristically answers to the AMS questions based on their differing levels of interest in physics as a subject.

As previously noted by Butcher and Jordan (2010), it is important to develop answer matching on the basis of real student responses which are gathered from a cohort that is essentially the same as those who are going to be the final users of an instrument. The FCI has a very varied user base, which means that the AMS (which is a free-response version of it) would most likely also inherit this variability. The findings from Q11, Q13 and Q19 have illustrated this effect in the context of a different user base, which was an important finding with respect to the design priorities of free-response format concept inventories. However, the findings also highlighted that these questions clearly required further development and testing.

Cases where the question was being used in the free-response format for the first time

In the cases of Q7, Q12, Q15, Q18, Q21, Q28 and Q33, the values for both the marking agreement and Cohen's kappa were outside the respective acceptable range of values. These seven questions were being used in the free-response format for the first time in Version 2 of the AMS, and their marking rules were inherited from similar questions from Version 1 of the AMS, as previously detailed in Table 7.1. Since the two questions in each pair are slightly different, it would be expected that false positive and false negative answers would occur, since the different situations can draw out different misconceptions and different types of wording for correct answers from the students. As previously, the false positives and false negatives were used to modify the marking rules on these questions; these new rules needed to be tested against further responses, since there could still be other ways in which students could answer the questions.

In the case of Q21, the false negative answers occurred because of an unforeseen problem in the question that arose from the diagram being changed from Version 1. The version of the question that was used in Version 2 of the AMS is shown in Figure

7.5 below. Students were asked to *"Identify the force or forces acting on the boy when he is at position P"*. Since the boy is clearly sitting on the swing seat in the diagram, answers such as *"Weight, Reaction from swing"* were recognized as being *correct* by human markers, but were not recognized as *correct* by the computer marking scheme, which had been finalized before the new diagram was inserted. The marking scheme was not adapted to account for this; instead, the diagram was changed to a *pendulum bob* scenario for a later version of the AMS, as shown in Figure 7.6 below. It was recognized that further testing of this new question and its associated marking scheme would be required in order to check that the issues had been resolved.

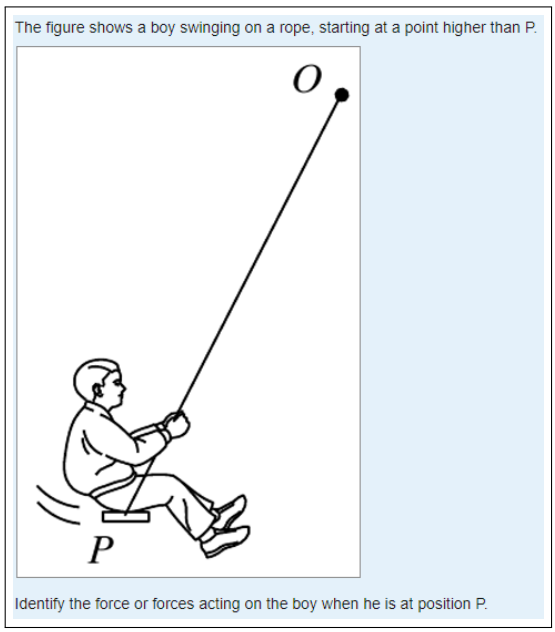


Figure 7.5: Q21 of Version 2 of the AMS, which is adapted from Q18 of the FCI.

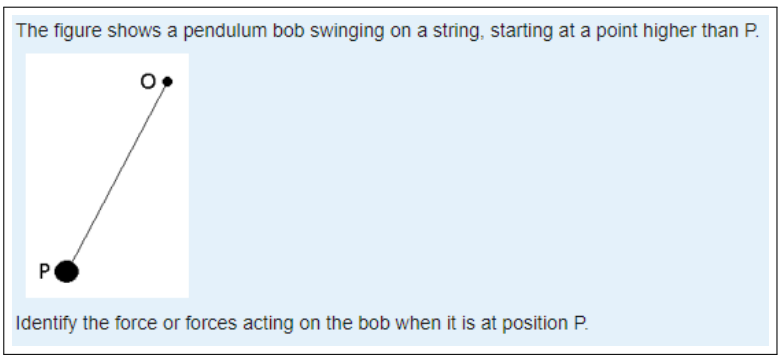


Figure 7.6: Q21 of Version 3 of the AMS, which is adapted from Q18 of the FCI.

7.4.2 Back-testing the Version 3 marking rules against the Version 2 responses

The Version 2 AMS computer marking rules were developed into the Version 3 AMS marking rules by making the changes discussed above, and these were used in the investigations described in **Chapter 8**. The Version 3 rules were subsequently back-tested against the Version 2 responses used to develop them, as a consistency check. The results of this back-testing are given in Table 7.6 below.

Question	Number of responses	Number of disagree-ments	Number of false positives	Number of false negatives	Marking agree-ment	Cohen's kappa
Q1	81	0	0	0	1.00	1.00
Q2	75	0	0	0	1.00	1.00
Q3	74	2	1	1	0.97	0.86
Q4	74	4	4	0	0.95	0.89
Q5	73	1	1	0	0.99	0.96
Q7(NEW)	71	0	0	0	1.00	1.00
Q11	66	0	0	0	1.00	1.00
Q12(NEW)	66	2	1	1	0.97	0.94
Q13	66	3	3	0	0.95	0.86
Q15(NEW)	66	5	3	2	0.92*	0.85
Q17	64	0	0	0	1.00	1.00
Q18(NEW)	64	4	2	2	0.94*	0.82
Q19	64	5	1	4	0.92*	0.84
Q20	64	1	1	0	0.98	0.96
Q21(NEW)	64	7	7	0	0.89*	0.77*
Q22	60	1	1	0	0.98	0.97
Q23	60	1	1	0	0.98	0.96
Q25	60	0	0	0	1.00	1.00
Q27	60	1	1	0	0.98	0.96
Q28(NEW)	60	0	0	0	1.00	1.00
Q29	60	2	2	0	0.97	0.91
Q30	60	1	0	1	0.98	0.96
Q31	60	0	0	0	1.00	1.00
Q32	60	5	4	1	0.92*	0.81
Q33(NEW)	60	4	2	2	0.93*	0.84

Table 7.6: Table showing the number of times the UHM disagreed with the Version 3 computer marking on the Version 2 free-response AMS questions and the nature of these disagreements, as well as the corresponding marking agreement and Cohen's kappa values for the UHM against the Version 3 computer marking.

Three different scenarios arose when considering the Version 2 IRR statistics and modifying the marking rules. In the first case, both the marking agreement and Cohen's kappa values were within the acceptable range, so the marking rules did not require much modification as a result. Q1, Q2, Q4, Q5, Q17, Q20, Q22, Q25, Q27, and Q31 were the questions in this scenario, and the results from these questions illustrate where automated marking schemes can be effective. Out of these questions, Q1, Q2 and Q17 had flawless IRR statistics, although it is worth bearing in mind that there were lower numbers of responses to Version 2 of the AMS than to Version 1.

In the second case, either one or both of the marking agreement and Cohen's kappa values were outside the accepted range of values. These questions had problems with the marking rules that could be resolved by considering the false positive and false negative answers to see where the rules were going wrong, and then modifying the marking rules based on these cases. Q3, Q7, Q11, Q12, Q13, Q15, Q18, Q19, Q23, Q28, Q29, Q30, Q32 and Q33 were the questions that made up this scenario, and the corresponding IRR statistics from the back-testing presented in Table 7.6 showed improvement for the performance of the marking rules on these questions. Further responses and IRR testing were required to check that these improved rules worked also on other responses.

The case of Q21 makes up the third case. For this question, the marking agreement and Cohen's kappa statistics had values that were outside the acceptable range, which highlighted a need to look again at the marking rules. However, unlike the other cases, the disagreement between the UHM and the computer marking was because of an error in the question diagram, which was subsequently resolved. More student responses were required to test whether the revised version of the question functioned normally. In summary, back-testing of the Version 3 marking rules against the Version 2 responses indicated that the revised marking rules were functioning well for every free-response question except Q21, although Q21 was found to have issues that were independent of its marking rules.

7.4.3 Back-testing the Version 3 marking rules against the Version 1 responses

As a further test for consistency, the Version 3 AMS marking rules were back-tested against the Version 1 UHM; the results of the IRR calculations comparing the Version 1 UHM to the Version 3 computer marking are shown in Table 7.7. Note that only 18 questions were in free-response format in the Version 1 AMS, so these are the only questions that have data for the Version 3 computer marking to be compared to.

Question	Marking agreement	Cohen's kappa
Q1	1.00	0.99
Q2	1.00	0.99
Q3	0.96	0.70*
Q4	0.96	0.91
Q5	0.99	0.98
Q11	0.96	0.92
Q13	0.97	0.94
Q17	0.98	0.96
Q19	0.95	0.89
Q20	0.99	0.96
Q22	0.99	0.97
Q23	0.94*	0.85
Q25	0.98	0.96
Q27	0.99	0.96
Q29	1.00	0.99
Q30	0.92*	0.76*
Q31	1.00	0.99
Q32	0.97	0.91

Table 7.7: Table showing the marking agreement and Cohen's kappa values for the Version 1 UHM against the Version 3 computer marking rules.

In the results given in Table 7.7, the values of both marking agreement and Cohen's kappa were within the acceptable range of values for questions Q1, Q2, Q4, Q5, Q11, Q13, Q17, Q19, Q20, Q22, Q25, Q27, Q29, Q31 and Q32. Because the Version 1 UHM was not used directly to build the Version 3 marking rules (although the overall development of the marking rules was an iterative process), this finding illustrates that the Version 3 marking rules were highly consistent for each of the 15 questions listed above. In the context of the aims of the overall study, this is an encouraging result for the development of automated marking schemes for these free-response AMS questions.

From the calculations presented in Table 7.7, three questions were identified as potentially being problematic; these were Q3, Q23 and Q30. For Q3, the marking agreement value was within the acceptable range, whereas the Cohen's kappa value was below the acceptable range of values. The issues with Q3 have been well documented across both the Version 1 and the Version 2 IRR studies, where the issues were found to arise from the fact that almost everyone who attempted Q3 got it right. Since Q3 did not appear on the original version of the FCI, it was flagged as a candidate for removal from the AMS, as issues with the automated marking of this question were persistent.

In the case of Q23, the marking agreement value was slightly below the threshold, but its Cohen's kappa value was within the acceptable range, and this implied that the marking rules were functioning in the intended way. As a result, there were no concerns about the automated marking of this question.

For the case of Q30, the values for both the marking agreement and Cohen's kappa were outside the acceptable range of values. Authoring marking rules for this question has been shown to be difficult over several versions of the AMS; this is because the correct answer from the original multiple-choice version contains two parts, *"slows down"* and *"stops"*, and it is difficult to capture the range of correct and incorrect answers that students give to this question using Pattern Match syntax. Since Q30 did appear on the original version of the FCI, it was flagged as a candidate to be reverted to its original multiple-choice variant, rather than for outright removal from the AMS.

7.4.4 Discussion of rule transfer between free-response questions

A facet of computer marking specific to the current study was the idea of transferring marking rules from one question to another in order to create new free-response questions. Seven selected-response questions from Version 1 of the AMS (Q7, Q12, Q15, Q18, Q21, Q28 and Q33) were replaced with free-response versions in Version 2 of the AMS. As a result, these questions inherited their marking rules from questions that tested related concepts, as previously detailed in Table 7.1 and its associated commentary.

In the cases of Q7, Q12, Q18 and Q28 of Version 2 of the AMS, transferring the rules appears to have been effective, with only small modifications being required to

effectively align the marking rules with the new question. In the cases of Q15 and Q33, the marking rules were inherited from the Q3 of Version 1 of the AMS, and it was found that several extra rules were required based on the different contexts of the questions. In the final case of Q21, an error with the question diagram meant that the responses did not match up with the intended marking rules, so it was not possible to draw any meaningful conclusions about the effectiveness of the rule transfer in this case. From the above considerations, it appears that transferring rules can be used to get some good initial marking rules, but these rules need to be further adapted based on the situation in order to function effectively.

Since the rules could not be directly transferred, the possibility of automating the rule creation process may be a more consistent approach to developing appropriate marking schemes for new free-response questions. Efforts have been made by Willis (2010) to apply the principles of machine learning to automatically generate mark schemes for Pattern Match questions, and applying this method to generating mark schemes for free-response AMS questions could be an avenue for further work.

7.4.5 Findings related to testing the human marking

Question	Marker 1 vs. UHM (MA)	Marker 1 vs. UHM (CK)	Marker 2 vs. UHM (MA)	Marker 2 vs. UHM (CK)	Marker 3 vs. UHM (MA)	Marker 3 vs. UHM (CK)	Marker 4 vs. UHM (MA)	Marker 4 vs. UHM (CK)	Marker 5 vs. UHM (MA)	Marker 5 vs. UHM (CK)
Q1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q2	1.00	1.00	0.99	0.97	1.00	1.00	0.99	0.97	1.00	1.00
Q3	0.99	0.93	0.99	0.93	0.96	0.78*	0.93*	0.52*	0.97	0.86
Q4	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.89	0.97	0.94
Q5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q7	1.00	1.00	1.00	1.00	0.99	0.90	0.96	0.75*	1.00	1.00
Q11	1.00	1.00	0.98	0.97	1.00	1.00	1.00	1.00	0.98	0.97
Q12	0.98	0.97	1.00	1.00	0.92*	0.85	0.97	0.94	0.98	0.97
Q13	1.00	1.00	0.98	0.95	0.97	0.90	0.98	0.95	1.00	1.00
Q15	0.98	0.97	0.95	0.91	0.86*	0.72*	0.92*	0.85	0.95	0.91
Q17	1.00	1.00	0.98	0.94	0.98	0.94	1.00	1.00	1.00	1.00
Q18	0.98	0.96	0.98	0.96	0.95	0.85	0.81*	0.56*	0.98	0.95

Table 7.8: Table showing the marking agreement (MA) and Cohen's kappa (CK) values of each human marker against the Version 2 UHM for the free-response AMS questions.

Question	Marker 1 vs. UHM (MA)	Marker 1 vs. UHM (CK)	Marker 2 vs. UHM (MA)	Marker 2 vs. UHM (CK)	Marker 3 vs. UHM (MA)	Marker 3 vs. UHM (CK)	Marker 4 vs. UHM (MA)	Marker 4 vs. UHM (CK)	Marker 5 vs. UHM (MA)	Marker 5 vs. UHM (CK)
Q19	0.94*	0.88	0.88*	0.75*	0.92*	0.84	0.72*	0.44*	0.75*	0.50*
Q20	1.00	1.00	0.98	0.96	1.00	1.00	0.98	0.96	1.00	1.00
Q21	0.88*	0.67*	0.97	0.93	0.97	0.93	0.81*	0.62*	0.78*	0.39*
Q22	0.95	0.90	0.98	0.97	0.98	0.97	1.00	1.00	1.00	1.00
Q23	0.98	0.96	0.98	0.96	0.95	0.88	1.00	1.00	1.00	1.00
Q25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q27	0.97	0.92	0.98	0.96	1.00	1.00	1.00	1.00	0.98	0.96
Q28	0.98	0.97	0.98	0.97	1.00	1.00	0.92*	0.83	0.97	0.93
Q29	1.00	1.00	0.98	0.96	0.95	0.87	0.98	0.96	0.90	0.72*
Q30	0.95	0.88	1.00	1.00	0.90*	0.75*	0.98	0.96	0.95	0.88
Q31	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.97	0.98	0.97
Q32	0.98	0.96	1.00	1.00	0.92*	0.82	0.95	0.88	0.93	0.85
Q33	0.98	0.96	1.00	1.00	0.98	0.96	0.95	0.89	1.00	1.00

Table 7.9: Table showing the marking agreement (MA) and Cohen's kappa (CK) values of each human marker against the Version 2 UHM for the free-response AMS questions.

The human markers were also checked for consistency. To do this, the marking agreement and Cohen's kappa statistics were calculated for each of the human markers against the Version 2 UHM. The results are presented in Tables 7.8 and 7.9 above. As was the case in Table 6.6, the abbreviation *MA* is used to denote *marking agreement*, and the abbreviation *CK* is used to denote *Cohen's Kappa*.

A high level of agreement was found between each of the human markers and the UHM on questions Q1, Q2, Q4, Q5, Q11, Q13, Q17, Q20, Q22, Q23, Q25, Q27, Q31 and Q33, since the marking agreement and Cohen's kappa values were within the acceptable range of values for these 13 questions. This in turn implied that the UHM was highly self-consistent for these questions.

Q12 and Q28 each had one instance where the marking agreement between human marker and UHM was below the acceptable value for marking agreement; whereas Q7 had one instance where the Cohen's kappa value was lower than the acceptable value. In addition, Q29 and Q30 each had one instance where the Cohen's kappa and marking agreement values between the human marker and UHM were both outside the acceptable range of values; and Q32 had two separate cases where the marking agreement between the human marker and the UHM was outside the acceptable range of values. However, the values were only slightly lower than the acceptable values in these cases, meaning that they could be attributed to the expected variation in the marking of the responses. As a result, the marking was not systematically different for the markers in these cases.

Q3 asked students to identify the forces acting on a stone after it has been dropped from a building, with the correct answer being *"weight"* (or equivalent). For this question, Markers 1, 2 and 5 each had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM. However, Marker 3 had an acceptable value for marking agreement, but their corresponding Cohen's kappa value was slightly below the acceptable value; as a result, this value could be attributed to expected variations in Marker 3's marking of the responses. In contrast, Marker 4 had values for marking agreement and Cohen's kappa that were outside the acceptable range of values. In contrast to the previous cases, the Cohen's kappa value was not close to the acceptable value, which indicated that there could be something different abut the way that Marker 4 marked the responses. It was found that Marker 4 was inconsistent when dealing with answers that contained both correct and incorrect information; for

instance, they marked the answer *"Air resistance, Gravitational Potential (Weight), Reaction Force"* as correct, whereas they marked the answer *"weight, downwards. Normal reaction force, upwards"* as incorrect, leading to the observed disagreement with the UHM.

Q15 required test-takers to identify the forces acting on a ball after it has been thrown upwards, with the correct answer being *"weight"* (or equivalent). Markers 1, 2 and 5 each had acceptable values for both the marking agreement and Cohen's kappa when compared with the UHM; whereas Marker 4 had an acceptable value for Cohen's kappa, but the corresponding marking agreement was below the acceptable range of values. In this case, the value was only slightly below the acceptable value, and it could be attributed to there being a smaller number of responses. However, Marker 3 had values for both marking agreement and Cohen's kappa which were outside the acceptable range of values, which implied that there was something different about the way that Marker 3 was marking the responses to this question. It was found that Marker 3 had consistently marked answers which added air resistance as incorrect, while the other markers had marked these answers as correct, leading to the observed discrepancy.

Q18 required students to compare the forces that a car and truck exerted on one another when the car was pushing the truck, with the correct answer being *"the forces are the same"* (or equivalent). For this question, Markers 1, 2, 3 and 5 all had acceptable values for both marking agreement and Cohen's kappa when compared with the UHM. However, Marker 4 had values for marking agreement and Cohen's kappa that were some distance below the acceptable values, which indicated that there was something different in the way that Marker 4 was marking this question. Examining the marking found that Marker 4 typically marked answers that used the word *"balanced"* as incorrect; the UHM instead marked these answers as correct, leading to the observed effect.

Q19 asked test-takers to to identify the forces acting on an elevator as it moved up an elevator shaft, with the sought forces being *weight* and *tension* (or equivalent). In the cases of Marker 1 and Marker 3, the Cohen's kappa values when compared with the UHM were acceptable, whereas the corresponding marking agreement values were outside the acceptable range of values. However, these values were only slightly below the acceptable value, so they could be attributed to there being a smaller number of

responses. However, in the cases of Marker 2, Marker 4 and Marker 5, the values of the marking agreement and Cohen's kappa were both below the respective acceptable values, and were sufficiently low to justify further investigation.

Marker 2 was found to be less generous than the UHM when it came to accepting alternative wordings for the *tension* force, as they did not accept wordings such as *"pulling force of cable"*. It is possible that Marker 2 interpreted the marking guidance in a different way than most of the other markers, leading to disagreement with the UHM. In a similar way, Marker 5 was very strict when marking the answers, only marking as correct those answers than mentioned that the *tension* was in the *cable*. As examples, the answer *"Weight and tension"* was marked as incorrect, whereas the answer *"Weight, tension in cable"* was marked as correct. Again, this behaviour did not match with that of the UHM, leading to the disagreement observed. On the other hand, Marker 4 had consistently accepted answers which contained additional forces, such as *"Gravity, tension, thrust"* as correct, whereas the other markers had not. It is possible that Marker 4 misinterpreted the intended marking guidance, leading them to mark any answer containing both *weight* and *tension* as correct.

Q21 asked test-takers to identify the forces acting on a boy while he is on a swing, with the required answers being *weight* and *tension* (or equivalents). For this question, Marker 2 and Marker 3 had acceptable values for the marking agreement and Cohen's kappa when compared to the UHM. However, in the cases of Marker 1, Marker 4 and Marker 5, the marking agreement and Cohen's kappa values were outside the acceptable range of values. In each case, these values were sufficiently far from the acceptable range to warrant further investigation. As detailed previously, a problem with the question diagram made it such that the situation did not match with the intended answer, and it appears that this adversely affected the consistency of the human markers, with each taking their own approach to the issue.

Marker 1 typically marked answers such as *"weight, reaction force"* as incorrect, while the UHM marked these answers as correct. This meant that Marker 1 adhered strictly to the marking guidance, despite having the option to use their own judgement to change their marking based on the new diagram. It is possible that Marker 1 did not notice the potential problem caused by the new diagram; or even that they felt as if students ought to model the *boy on a swing* as a *particle on a string*, and give an answer based on that interpretation. Further, Marker 5 was found to have been very

strict in deciding what was required for the answer to be marked as correct, accepting only the small number of answers that mentioned that the *tension* force was in the *rope*; this behaviour was not observed with the other markers. In contrast, Marker 4 consistently marked answers such as *"Weight, centripetal force"* as correct, while the UHM typically marked these answers as incorrect. However, for cases such as *"weight, reaction force"*, Marker 4 sided with the UHM and marked the responses as correct. This shows that Marker 4 used their own interpretation of the marking guidance when marking this question, and this serves as an illustration of where human markers can be inconsistent.

From the above considerations, two questions were particularly problematic for the human markers; these were Q19 and Q21. In both of these cases, the marking agreement and Cohen's kappa values between various human markers and the UHM were not high. The issue with Q19 was found to arise because different markers applied different levels of strictness when accepting alternative wordings for the *tension* force, and there appears to be some level of subjectivity involved in marking this question as a result. The issue with Q21 occurred because of an error with the question diagram; different markers reacted in a variety of ways to this, with some altering their marking to match the new situation, and others not doing so. Since the responses did not match up with the intended question, human marking of Q21 was highlighted as being a particularly subjective process.

The answers to the free-response AMS questions were marked by 5 expert human markers, and the UHM was constructed by taking the *majority view* as the mark awarded for each response. *Borderline* cases arose when 3 markers chose to mark the responses in one way, and the other 2 markers chose to mark the same response in the other way. For triangulation with the above findings, the number of these borderline cases encountered in the human marking of each free-response question on Version 2 of the AMS are given in Table 7.10 below. Table 7.10 shows that most of the questions had a small number of borderline cases, although this may be expected since there were fewer responses to the AMS questions in the Version 2 study. The only question with a high number of borderline cases was Q21; this result agrees with the findings from the above human marking study, where marking issues were highlighted as a result of this question not matching with its intended marking scheme. As previously detailed, Q21 was modified to align the question being asked with the marking scheme, although this still required further testing to check whether it has successfully resolved the issues.

Version 1 AMS free-response question	Number of responses	Number of borderline human marking cases
Q1	81	0
Q2	75	1
Q3	74	4
Q4	74	0
Q5	73	0
Q7	71	1
Q11	66	0
Q12	66	2
Q13	66	1
Q15	66	3
Q17	64	1
Q18	64	4
Q19	64	3
Q20	64	0
Q21	64	11
Q22	60	0
Q23	60	0
Q25	60	0
Q27	60	1
Q28	60	1
Q29	60	3
Q30	60	4
Q31	60	0
Q32	60	3
Q33	60	1

Table 7.10: Table showing the number of borderline cases encountered in the human marking of each of the free-response questions on Version 2 of the AMS.

7.4.6 Discussion of human and computer marking

Across the Version 1 and Version 2 IRR studies, there were examples that highlighted the strengths and weaknesses of both human and computer marking. Q30 of the AMS caused such problems for the computer marking and the human markers in the Version 1 IRR study. The question was adapted from Q27 of the FCI, and it requires test-takers to describe what happens to the speed of a box after the force being applied to it is removed. The correct answer to this question has two parts, *"slows down"* and *"stops"*. Answers which only referred to the box *"slowing down"* were also considered acceptable, while answers which referred to the box stopping immediately (such as *"it stops"*) were not. These marking conditions were difficult to program into Pattern Match rules for the computer to follow, which led to the computer marking being less

effective on this question. In addition, although the marking guidance provided to the human markers gave examples of common *correct* and *incorrect* answers, these were necessarily not exhaustive, so there were instances where the markers had used their own judgement. On some occasions this led to a lack of consistency in the human marking.

Q21 of Version 2 of the AMS asked test-takers to identify the forces acting on a boy while he was in motion on a swing. Issues arose with the question because the diagram accompanying the question represented a different situation from the question being asked. As a result of this, some of the responses to the question answered the *intended question*, whereas others answered the *question being asked by the diagram*. This caused problems for both human and computer marking. For the human marking, different human markers reacted differently to the responses, with some sticking to the marking guidance and marking only those answers which correctly answered the *intended question* as correct; and others adapting to the situation by also marking answers that correctly answered the *diagram question* as correct. For the computer marking, the marking rules were set up to only mark correct answers to the *intended question* as correct, with the computer having no awareness of the issues thrown up by the mis-matched diagram.

Taken together, the issues with Q30 and Q21 illustrate instances where the greatest strengths of human and computer marking were turned against them. For the human marking, the human marker's ability to interpret the question and marking guidance led to different markers awarding different marks, which highlighted the subjective nature of human marking in some cases. For the computer marking, being totally objective meant that some *correct* answers were marked as *incorrect* because they did not line up with the programmed marking rules. However, the human and computer markers were ultimately not to blame for the marking issues in these questions; the question wording and design was faulty, meaning that the markers could never match up against the intended marking scheme. In these cases, the fault lies entirely with the question designers, and it once again returns to the important point that question design and wording needs to be carefully considered when trying to improve the performance of an automated marking scheme (Butcher and Jordan, 2010).

Another common issue that arose for human and computer marking were answers where students had included both correct and incorrect information in the answer,

although the problems with marking such answers were different between human and computer markers. Since no partial credit was available, these answers were marked as either correct and incorrect, and human markers are known to be inconsistent when handling such cases (Butcher and Jordan, 2010); as such, these answers highlighted concerns about the consistency of human marking. Computer marking also struggled with such cases, in a large part because the computer does not have a sufficiently advanced appreciation of semantics, meaning that the computer is incapable of offering *benefit of the doubt* or other subjective judgement. This finding agrees with the previous reflections of Mitchell et al. (2002), who identified that marking responses containing both correct and incorrect parts could prove to be a serious problem for the advancement of automated free-response marking.

The above findings indicate that computer marking would be expected to outperform human marking whenever consistency and/or objectivity is the dominant requirement of the marker. However, the Version 1 and Version 2 IRR studies also highlighted some instances where human marking would always outperform computer marking; these arose in Q19 of Version 1 of the AMS, and in Q3, Q29 and Q31 of Version 2 of the AMS. In these cases, a small number of students had given answers using precise or unorthodox wording. For human markers, this wording could be recognized and marked appropriately using past experience and judgement; in contrast, it was not possible to have the computer marking rules cover these cases without *over-fitting* to account for them. As a result, these findings highlight that human marking would be expected to outperform computer marking whenever subjective judgement is required by the marker.

7.5 Limitations of the data collected

Several points for discussion were raised by the properties of the data used for the CTT and IRR testing in both **Chapter 6** and **Chapter 7**. The AMS questions were marked using a discrete marking metric, meaning that responses were marked as either correct or incorrect, with no partial credit given. It follows that the corresponding calculations gave an output which contained a single number, with no further measure of confidence given for the result; this is acknowledged as a limitation of the study. In future studies, this effect could potentially be mitigated with the use of marking schemes which make use of continuous marking metrics (these would allow partial credit to be given), although this was beyond the scope of the current study.

Version 1 of the AMS had $N = 254$ competed attempts, whereas Version 2 of the AMS had $N = 60$ completed attempts; this meant that a smaller sample size was used in the Version 2 study. Since CTT is grounded in the idea of finding out whether tests and their questions are reliable, it follows that having more completed attempts to analyze is preferable. This means that on a rough qualitative level, the CTT calculations for Version 1 of the AMS should be more reliable than those for Version 2. For the IRR statistics, the effects of having a smaller sample size were more immediately obvious. When there was a smaller number of responses in Version 2, the IRR statistics were more sensitive to small numbers of disagreement cases, leading to some of the results being skewed towards the higher end (and others to the lower end) of the IRR statistics' ranges. This effect was countered in the interpretation of the results by looking at how many false positive and false negative cases arose, and deciding whether or not the marking rules had systematic failings based on both the IRR statistics and observations from the responses given by students. In addition, more effective marking rules can be developed when more responses are used in building the answer matching (Butcher and Jordan, 2010). This point was highlighted by the underperformance of the marking rules in some of the free-response questions on Version 2 of the AMS, since these were developed with a smaller number of responses.

The Version 1 and Version 2 cohorts were characteristically different. The Version 1 sample was drawn from a population of high school physics students, OU undergraduate students from various modules, and undergraduate physics students from a university external to the OU; most of these students would be expected to have encountered Newtonian mechanics before attempting Version 1 of the AMS, meaning that they would be expected to attain higher scores on the AMS questions. In contrast, the Version 2 sample was drawn from a population that contained mostly OU undergraduate students on the level 1 module *S112 Science: Concepts and practice*. These students would be expected to have had less previous exposure to Newtonian mechanics, because these students are not likely to be registered on physics or astronomy pathways, and would be expected to attain lower scores on the AMS questions. Indeed, the mean score on Version 1 of the AMS was higher than that on Version 2, and comparing the results from the difficulty calculations from both cohorts also leads to the conclusion that the AMS questions were harder for the Version 2 cohort. These findings indicate that there was a consistent and noticeable cohort effect, and this arose because of the differences in Newtonian mechanics study experience between

the Version 1 and Version 2 cohorts.

The AMS was adapted from the FCI, which means that it has a wide range of potential users with different levels of prior study experience and understanding of physics; it was therefore useful to test the AMS with cohorts which were characteristically different. In addition, the responses used for marking rule development need to be representative of the potential user base, which further highlights why testing the AMS with a wide range of users is an important consideration in the development process.

Version 1 and Version 2 of the AMS were both administered as optional activities with no additional incentive, meaning that any participants who worked through the tests and completed all of the questions did so voluntarily. It is hence probable that only the most capable and enthusiastic students completed the tests (Hunt and Jordan, 2016), meaning that the scores obtained and analyzed were likely higher than would have been expected from a more *typical* cohort. This potentially affected the calculation of the various CTT and IRR statistics. For example, since students on average gave more correct answers, it follows that the difficulty values of the questions were likely to be over-estimated, because the students found the questions to be easier. Additionally, because the students were higher-performing on average, it also follows that the discrimination values were likely to be under-estimated, because there was not a wide range of abilities to differentiate between in the first place.

Ideally, this self-selection effect could be avoided in future studies by making completion of the tests into mandatory exercises, and by having the students complete them at a certain time. However, this approach is not straightforward in a study that mainly makes use of Open University students in the data gathering step; this is because The Open University is a distance learning institution where there are no set lectures, and students flexibly work through course material in their own time. In addition, most Open University students study part-time, so have a variety of other pressures on their time. To counteract this self-selection effect, a cohort including other settings would be required. This idea formed the basis of the study conducted with Version 3 of the AMS, which is covered in **Chapter 8**. In this study, the cohort consisted of mainly high school students, with data gathered through a different platform than the OSL.

7.6 Conclusions

The primary objective of the study presented in this chapter was to further test the reliability of the AMS questions and marking rules. To achieve this, Version 2 AMS response data were collected from students during the academic year 2018-2019. To test the AMS questions for reliability, *Classical Test Theory* (CTT) statistics were calculated on the Version 2 data set, and it was found that the Version 2 AMS questions mostly functioned well. Taking this together with the previous CTT findings pertaining to the Version 1 AMS questions, this meant that beyond the well-documented case of Q3, the AMS questions were stable and did not require further development.

To test the AMS marking rules for reliability, *Inter-Rater Reliability* (IRR) statistics were calculated on the Version 2 data set. There were 25 free-response questions to test in the Version 2 IRR study. Of these, seven were adapted from selected-response questions from Version 1 of the AMS, meaning that these questions were being used in free-response format for the first time; hence the computer marking was not expected to be highly functional in these cases. For the other 18 free-response questions on Version 2 of the AMS, results from the IRR studies indicated that the computer marking was performing well on half of the questions; this indicated that the computer marking rules still required further development.

The secondary aim of the IRR study was to investigate the effectiveness of transferring rules between free-response questions which tested similar concepts and content. It was found that rule transfer could not be effectively done as a direct one-to-one mapping between paired questions, because the different contexts presented in each of the questions led to students giving sufficiently different answers. However, it was found that the rules could be made to function effectively after modifying them based on the false positives and false negatives brought up by the different contexts in the questions, although further testing would still be needed. The idea of automating the rule creation process to improve consistency was identified as a possible future direction for the research.

The findings from the CTT and IRR studies conducted in this chapter were used to iterate Version 2 of the AMS into Version 3. This completed another step in the development process of the AMS. In order to check the level of functionality of the computer marking, the Version 3 AMS marking rules needed to be tested for reliability; this testing is the focus of the next chapter.

7.7 Summary and looking ahead

Chapter 7 presented the quantitative findings from the *Classical Test Theory* (CTT) and *Inter-Rater Reliability* (IRR) studies conducted using responses gathered to Version 2 of the AMS. The findings from the CTT strand of the study indicated that the AMS questions were mostly functioning well, and had stabilized with respect to the previous version; whereas the findings from the IRR strand of the study showed that the corresponding AMS marking rules still required further development.

Chapter 8 focuses on further efforts to develop and test the AMS marking rules in a new setting. It presents findings from data gathered by administration of the AMS through the *Isaac Physics* platform in the academic year 2019-2020.

8 Applying the Alternative Mechanics Survey in a wider educational context

8.1 Rationale

In **Chapter 7**, the development and testing of the AMS questions and marking rules using Version 2 AMS data collected in the academic year 2018-2019 was presented. The findings from the *Classical Test Theory* (CTT) strand of the study showed that the questions functioned well overall. However, the findings from the *Inter-Rater Reliability* (IRR) strand of the study showed that the corresponding AMS marking rules still required further development and testing in order to reach the required level of functionality; this is one of the objectives of the work presented in the current chapter. The AMS was developed and tested at The Open University through the Moodle *OpenScience Laboratory* (OSL) platform. Another of the aims of the overall research was to develop conceptual evaluation tools that could be widely used, and the potential usefulness of the AMS would be enhanced if it could be made available on a range of different platforms. The second objective of the work presented in the current chapter was to extend the use of the AMS to a wider context by administering it on a different educational platform with a different user-base, *Isaac Physics*.

8.2 Methods

8.2.1 Data collection

The Version 3 AMS data set was collected in the academic year 2019-2020 by splitting the AMS into three shorter length tests, each designed to have similar conceptual balance to the overall AMS. This data-gather approach was similar to the one employed by Han et al. (2015; 2016), who split the original FCI into two half-length tests. The AMS questions used in each test are shown in Tables 8.1, 8.2 and 8.3 below. Note that the standardized AMS question numbering is used in these tables, and throughout the remainder of this chapter. Further, the abbreviation FRQ is used for free-response questions that require a short phrase or sentence to answer, and the abbreviation FRQ(L) is used for free-response questions that require a single-letter answer.

Question	Question type	Theme	Concept	AMS question
1	FRQ	Car collision	Newton's Third Law	Q5
2	FRQ	Stone drop	Newton's Second Law	Q3
3	FRQ	Stone drop	Newton's First Law	Q4
4	FRQ	Marble in track	Newton's Third Law	Q7
5	FRQ(L)	Marble in track	Newton's First Law	Q8
6	FRQ(L)	Cannon	Newton's Second Law	Q14
7	FRQ	Boy on swing	Newton's Second Law	Q21
8	FRQ	Woman pushing box	Newton's First Law	Q28
9	FRQ	Woman pushing box	Newton's Second Law	Q29
10	FRQ	Woman pushing box	Newton's First Law	Q30
11	FRQ	Office chairs	Newton's First Law	Q31

Table 8.1: Table showing which AMS questions were used to assemble Isaac Test 1.

Question	Question type	Theme	Concept	AMS question
1	FRQ	Ball toss	Newton's Second Law	Q15
2	FRQ	Balls on table	Newton's Second Law	Q1
3	FRQ	Balls on table	Newton's Second Law	Q2
4	FRQ(L)	Bowling ball	Newton's First Law	Q16
5	FRQ	Elevator	Newton's First Law	Q19
6	FRQ(L)	Hockey	Newton's First Law	Q10
7	FRQ	Hockey	Newton's First Law	Q11
8	FRQ	Hockey	Newton's First Law	Q12
9	FRQ	Hockey	Newton's Third Law	Q13
10	FRQ	Truck and car	Newton's Third Law	Q17

Table 8.2: Table showing which AMS questions were used to assemble Isaac Test 2.

Question	Question type	Theme	Concept	AMS question
1	FRQ	Office chairs	Newton's Third Law	Q32
2	FRQ	Moving blocks	Newton's Second Law	Q22
3	FRQ	Moving blocks	Newton's Second Law	Q23
4	FRQ	Elevator	Newton's First Law	Q20
5	FRQ	Truck and car	Newton's Third Law	Q18
6	FRQ(L)	Rocket	Newton's Second Law	Q24
7	FRQ	Rocket	Newton's Second law	Q25
8	FRQ(L)	Rocket	Newton's First Law	Q26
9	FRQ	Rocket	Newton's First Law	Q27
10	FRQ(L)	Hammer throw	Newton's First Law	Q8
11	FRQ	Tennis player	Newton's Second Law	Q33

Table 8.3: Table showing which AMS questions were used to assemble Isaac Test 3.

These tests were put onto the University of Cambridge's *Isaac Physics* platform, which hosts physics activities and tests at various education levels. Data were collected by offering the AMS tests as an activity on the Isaac Physics site, which has a primary user base of high school students, undergraduate university students and high school teachers; this differs from the main user base of the OSL, which is almost exclusively Open University undergraduate students. In another difference from the OSL versions of the AMS, the Isaac Physics questions told test-takers whether their typed answer was marked as right or wrong by the computer in real-time, thus providing them with some instantaneous, basic-level feedback on their performance. Once users had completed the tests, the data were downloaded from the Isaac Physics site, where it had been marked by the Version 3 AMS marking rules. All non-blank entries for each question were retained for calculation of the IRR statistics.

Since a thorough investigation into human marking had already been completed (See **Chapter 6** and **Chapter 7**), and the response matching was further advanced, the number of human markers was reduced from five to three. Initial marking was done by the author and one member of the supervisory team, with a second member of the supervisory team arbitrating when the others disagreed, in order to establish the *Unified Human Marker* (UHM) for use in comparison with the computer marking. Both supervisors had been amongst the previous markers, so were very familiar with the AMS questions and marking guidelines.

8.2.2 Data analysis

The objective of the study was to test and develop AMS marking rules, not AMS questions, so no CTT analysis took place here. To test the marking rules, the IRR statistics of marking agreement and Cohen's kappa were used, as explained in **Subsection 6.2.2.** and used throughout **Chapter 6** and **Chapter 7**; they were used in the current study as follows.

For each of the free-response questions, the marking agreement and Cohen's kappa statistics were calculated for the UHM against the computer marker; the values of these were then used to identify any problematic cases where the computer marking rules were not functioning at the required level. The number of times that the UHM disagreed with the computer marker on each question was counted, and the number of *false positives* and *false negatives* were also counted. The false positives and false

negatives were used to develop new marking rules.

After making suitable changes to the marking rules, the versions of the marking rules pre-change and post-change were further tested against the UHMs from previous academic years, by calculating the marking agreement and Cohen's kappa statistics for these UHMs against the computer marking rules. This extra testing was carried out in order to check for consistency between different iterations of the marking rules. The Version 3 AMS questions used to conduct these studies can be found in **Appendix E**.

8.3 Results and Discussion: AMS Version 3 IRR study

8.3.1 Marking agreement and Cohen's kappa

The changes suggested to the Version 2 marking rules in **Chapter 7** were implemented to develop the Version 3 marking rules. Hence, the marking agreement and Cohen's kappa values were calculated for the Version 3 computer marking against the Version 3 UHM marking of the AMS free-response questions. The results are shown in Table 8.4, along with information pertaining to the number of times that the UHM and computer marker disagreed. At this point, it is worth recalling that the acceptable range of values for marking agreement are [0.95, 1], whereas the acceptable range of values for Cohen's kappa are [0.8, 1]. With these ranges in mind, the different scenarios that emerged from the results presented in Table 8.4 are discussed afterwards.

Question	Number of responses	Number of disagree-ments	Number of false positives	Number of false negatives	Marking agree-ment	Cohen's kappa
Q1	45	0	0	0	1.00	1.00
Q2	44	1	1	0	0.98	0.95
Q3	107	3	3	0	0.97	0.90
Q4	105	2	2	0	0.98	0.93
Q5	118	1	0	1	0.99	0.98
Q7	99	0	0	0	1.00	1.00
Q11	39	0	0	0	1.00	1.00
Q12	41	0	0	0	1.00	1.00
Q13	40	0	0	0	1.00	1.00
Q15	47	3	3	0	0.94*	0.84
Q17	40	0	0	0	1.00	1.00
Q18	29	0	0	0	1.00	1.00
Q19	32	0	0	0	1.00	1.00
Q20	29	0	0	0	1.00	1.00
Q21	96	5	5	0	0.95	0.88
Q22	31	1	1	0	0.97	0.93
Q23	23	3	2	1	0.87*	0.73*
Q25	29	0	0	0	1.00	1.00
Q27	30	0	0	0	1.00	1.00
Q28	91	4	1	3	0.96	0.88
Q29	92	0	0	0	1.00	1.00
Q30	92	7	1	6	0.92*	0.79*
Q31	87	0	0	0	1.00	1.00
Q32	36	0	0	0	1.00	1.00
Q33	32	2	2	0	0.94*	0.82

Table 8.4: Table showing the number of times the UHM disagreed with the Version 3 computer marking on the Version 3 free-response AMS questions and the nature of these disagreements, as well as the corresponding marking agreement and Cohen's kappa values for the UHM against the Version 3 computer marking.

Cases with acceptable values for both marking agreement and Cohen's kappa

For the cases of Q1, Q2, Q3, Q4, Q5, Q7, Q11, Q12, Q13, Q17, Q18, Q19, Q20, Q21, Q22, Q25, Q27, Q28, Q29, Q31 and Q32, each question had acceptable values for both the marking agreement and Cohen's kappa, so there were no concerns about the functionality of the marking rules. In addition, Q1, Q7, Q11, Q12, Q13, Q17, Q18, Q19, Q20, Q25, Q27, Q29, Q31 and Q32 each had no disagreement cases between the UHM and the computer marker, so the marking rules were functioning at a particularly high level for these questions. This was a particularly pleasing result because the user base was different from that used in the Version 1 and Version 2 studies, and it was

recognized that they might therefore have given characteristically different responses. Although the sample size was rather small for the Version 3 study, the positive result gave encouragement that the rules were now functioning well for a range of users.

For most of the other questions, the false positive and false negative cases that arose were used successfully to develop the marking rules, using the same approach as previously (see **Subsection 6.2.2**. and **Subsection 6.4.2**). In the case of Q4, it was found that adding extra marking rules to cancel out the false positive cases led to the creation of new false negative cases, which made improving the marking rules difficult. In the context of the current work, avoiding false negatives was seen as a priority over avoiding false positives, since students should not be led to believe that they misunderstand a topic when they actually do understand it; this is particularly important when students are being given immediate feedback on whether their response is correct or not. However, in the cases of Q5 and Q28, it was found that some of the correct answers were given using unconventional phraseology. As a result, there was no straightforward way of adding new marking rules to account for these cases without resorting to *over-fitting* (Zehner et al., 2016); this illustrates cases where it is difficult to avoid false negatives.

Discussion of the case of Q21

Q21 asked test-takers to identify the forces acting on a boy while he was in motion on a swing, and it had previously been problematic for both human and computer marking in Version 2 of the AMS (see **Subsection 7.4.1** and **Subsection 7.4.5**), because the diagram provided caused the question and the marking rules to not match up. Because of the timescales of the collaboration with Isaac Physics, it was not possible to update the diagram in this question based on this finding before it went live on the platform. In spite of this, Q21 had acceptable values for both marking agreement and Cohen's kappa, and these values were much improved from the Version 2 testing. As a result, Q21 was not identified as being a problematic case in the Version 3 IRR study. This outcome can be explained as follows.

The marking rules for Q21 were designed to accept *tension* and *weight* as the two correct forces acting on the boy, which is referred to as the *first interpretation* in what follows. However, the diagram itself depicts a situation where *normal reaction force* and *weight* could also be accepted as the correct answers, and this is known as the

second interpretation in what follows. This means that the question is open to at least two different interpretations, with a set of correct answers associated with each interpretation.

For the Version 2 AMS cohort, many of the students gave answers using the second interpretation, such as *"weight and normal reaction force"*, and these were marked as incorrect by the computer, leading to a large number of disagreement cases with the UHM. For the Version 3 cohort, most of the students gave answers using the first interpretation, such as *"weight and tension"*, and these were marked as correct by the computer. It follows that students did not give incorrect answers because they used the second interpretation; rather, the majority of the incorrect answers displayed some sort of misunderstanding pertaining to the situation. Common incorrect answers missed out one of the two forces, or incorrectly identified *energy* as a force. The latter of these misunderstandings was not represented in any of the selected-response versions of Q21, and this is an example of the selected-response question format being unable to adequately assess student knowledge, which is a point that has previously been raised in the literature by Dufresne et al. (2002).

There is also a possible cohort effect associated with Q21. Some of the test-takers from the Version 2 cohort used the first interpretation to answer the question, whereas others used the second interpretation; this contrasts with the situation observed for the Version 3 cohort, where most of the students used the first interpretation. The first interpretation requires the test-taker to model the situation as a pendulum bob, and it is possible that the Version 3 cohort had more previous exposure to this sort of problem-solving technique than the Version 2 cohort. Since the first interpretation is the intended question, the diagram for Q21 was changed to that of a pendulum bob in the final version of the AMS, with the aim of scaffolding test-takers towards using the first interpretation to answer the question.

Cases with acceptable values for Cohen's kappa, but lower values for marking agreement

For Q15 and Q33, the values for marking agreement were just below the acceptable threshold, but the Cohen's kappa was within the acceptable range, and the values of both statistics had improved since the question was first tested on the Version 2 cohort (see Table 7.5). All the cases of disagreement between the computer marking and

the UHM were false positive cases where students had added incorrect information to their otherwise correct answers, and the marking rules were modified using these cases without difficulty. Overall, there were therefore no serious concerns with the functionality of the marking rules for these questions, although further testing was still needed.

Cases with lower values for marking agreement and Cohen's kappa

The values for the marking agreement and Cohen's kappa were both outside the range of acceptable values for Q23, which highlighted that there could be problems with the Version 3 marking rules for this question. AMS Q23 corresponds to Q20 of the original FCI, and it asks test-takers to identify whether either of a moving pair of blocks is accelerating, with the sought correct answer being *"neither"* (or equivalent). On Q23, there were 3 cases where the UHM disagreed with the human marker. Of these, 2 cases were false positives, with the other case being a false negative. The false positive cases occurred because the respondents had given incorrect answers that satisfied the *correct* computer marking criteria, such as *"block a block b are both accelerating"*, and these cases were handled by adding extra rules to negate on the incorrect information in the answers. The false negative case occurred because the respondent had used a specific wording to give the correct answer, *"A not, B not"*, and there was no straightforward way to write a marking rule to cover this case, hence it was left as is.

From the above examples, issues with the Q23 marking rules arose because test-takers gave answers that referred to the two blocks *"Block A"* and *"Block B"* separately, meaning that these accelerations were discussed separately in the answers. This issue does not arise on the original FCI version of the question because the multiple-choice options are designed to *complete* the narrative that is started in the question statement. As a result, the AMS question wording was modified (see Figure A.41 in **Appendix A**) to discourage students from referring to the two blocks separately.

Issues were previously found with the computer marking of Q23 in the Version 1 IRR study. However, the situation was improved in the Version 2 IRR study, where it was found that Q23 was not problematic for the computer to mark. Since issues with the computer marking of Q23 arose again in the Version 3 IRR study, an explanation is required for the varying effectiveness of the marking rules. One possible explanation

is that the different number of responses used in each of the IRR studies caused the variability of the calculated IRR statistics from each study. In particular, it is possible that the smaller number of responses inflated the negative effects of the disagreement cases for the Version 3 results. Another possible explanation is that the different values of the IRR statistics from each study were the result of cohort effects. The cohorts used to conduct the IRR testing in each academic year had different previous exposure to the material, and would be expected to give different types of answers based on how they had been taught. It is likely that both factors contributed in some way to the different values of the IRR statistics.

Q23 was a question where several students abandoned their AMS attempts in the Version 1 study. Since the AMS was administered as three sub-tests in the Version 2 study, there is no data to make a comparison of this effect to in the current study. However, the previous drop-off in participation was not likely to be the result of a computer marking effect, since students were not told whether their answers were right or wrong in Version 1 of the AMS. It is instead possible that this drop-off was caused by students having to read too much on-screen to answer the questions, an effect previously reported by Nardi and Ranieri (2019). Having three sub-tests may have mitigated this effect for Version 3 of the AMS, although this is not measurable without further supporting data.

Q30 had values for both the marking agreement and Cohen's kappa that were outside the acceptable range of values, so there could be issues with the Version 3 marking rules. It corresponds to Q27 of the original FCI, and the question asks test-takers to describe what happens to the speed of a box after the force being applied to it is removed, with the sought correct answer being that it *slows down and stops* (or equivalent). Overall, the UHM disagreed with the computer marker in 7 instances. Of these, 1 case was a false positive, and the other 6 cases were false negatives. The false positive case arose because the respondent had given a subjective answer that the UHM voted to mark as incorrect (the response went to the third marker in this case), *"speed of the box comes to rest"*, so no changes were made to the marking rules based on this case. The false negative cases occurred because respondents had given correct answers to the question in ways that were not recognized by the marking rules, for example *"it will slide slightly then stop"*, and these cases were handled by adding extra marking rules to include the cases.

Q30 was previously identified as problematic for the human and computer markers in the Version 1 AMS study, and this behaviour has re-occurred in the current study of Version 3 of the AMS. In both studies, transferring both parts of the model correct answer to this question (*slows down* and *stops*) into corresponding Pattern Match marking rules was found to be difficult, since several answers contained only one part of it. Furthermore, different human markers applied different degrees of strictness when marking the responses to this question, based on their own interpretations of the marking guidance. The repeated issues identified when marking this question pointed to there being flaws in the question itself, and such issues have previously been raised in the literature.

Rebello and Zollman (2004) noted that the original wording for FCI Q27 did not mention *friction* explicitly, although *friction* is required in order to answer the question in the desired way. They also pointed out that students can be accustomed to dealing with *frictionless* surfaces, so they may answer a different question from the one that is intended. This echoes the problems that the human and computer marking had with this question, as it is difficult to come up with a consistent marking scheme when the question (and what it is asking for) is itself unclear. For instance, if a student interprets the question as asking about a *smooth* surface, then an answer that states that *the box continues to move* is not wrong; or if a student interprets the floor as being very *rough*, then an answer that states that *the box stops immediately* is not wrong.

From the above considerations, the issues with the computer marking of Q30 were found to arise from the unclear wording of the question, and the ambiguous types of answers that the students gave in response to this. As a result, the largest barrier to making Q30 into an effective free-response question appears to be that the difference between correct and incorrect answers to the question is not clearly defined, which is a crucial design priority for automated marking to be effective (Jordan and Mitchell, 2009). Since there was no obvious way of rectifying this without making significant changes to the question, Q30 was reverted to multiple-choice format for the final version of the AMS.

Summary

Four different cases arose when considering the IRR statistics and improving the marking rules. In the first case, both the marking agreement and the Cohen's kappa values were within the acceptable range, and the marking rules were not modified. Q1, Q2, Q4, Q5, Q7, Q11, Q12, Q13, Q17, Q18, Q19, Q20, Q21, Q25, Q27, Q28, Q29, Q31, and Q32 were in this scenario, which is a very encouraging result for the development of the automated marking schemes for the free-response AMS questions. Out of these cases, Q1, Q7, Q11, Q12, Q13, Q17, Q18, Q19, Q25, Q27, Q29, Q31 and Q32 had flawless IRR statistics, although it is worth bearing in mind that some of these questions had smaller numbers of overall responses than others, meaning that further testing is needed.

In the second case, both the marking agreement and the Cohen's kappa values were within the acceptable range, but the marking rules needed to be slightly modified based on the disagreement cases; Q3 and Q22 were the questions in this scenario. In the third case, the Cohen's kappa value was acceptable, but the marking agreement was slightly below the acceptable value. In the same way as for the second case, the marking rules needed to be slightly modified in these cases; Q15 and Q33 were the questions in this scenario.

Questions Q23 and Q30 make up the fourth case. For these questions, the marking agreement and Cohen's kappa statistics had values that were outside the acceptable range, which highlighted a need to look again at the marking rules. In the case of Q23, rules to negate on incorrect information were added to counter the false positive cases and the question wording was altered; whereas in the case of Q30, it was concluded that the question would be more effective if asked in the multiple-choice format.

8.3.2 Establishing the final version of the AMS for this project

Implementing the changes to the Version 3 marking rules based on the above IRR calculations led to the final version of the marking rules for each of the free-response questions. Based on the findings from the studies conducted in **Chapter 6**, **Chapter 7** and the current chapter, changes were made to the wording of three of the questions in the final version of the AMS; two questions changed format for the final version of the AMS; two questions were removed from the final version of the AMS; and one question was restored to the final version of the AMS from a previous version. These

changes are explained below, with the standardized AMS question numbering system used throughout to avoid ambiguity.

Cases where the question wording was changed

Q11 asks test-takers to compare the speeds of a hockey puck before and after it has been kicked, and it is adapted from the situation presented in Q9 of the FCI. In the question setup, speeds u and v are used to denote the *horizontal* and *vertical* components of the resultant velocity respectively. However, there was anxiety that students might confuse u and v respectively with the *initial* and *final* speeds of the puck in their answers. As a result, Q11 had its wording changed such that the speeds u and v defined in the question are denoted respectively instead as r and k, with the aim of making the question wording more straightforward to interpret.

Q21 of the AMS corresponds to Q18 of the FCI, and it asks test-takers to identify the forces acting on a boy while he is on a swing. For the case of Q21, the question was changed from the situation involving a *boy on a swing* to a situation involving a *pendulum bob* in the final version of the AMS, since the diagram accompanying the boy on a swing scenario added extra forces that the question was not designed to ask about. The issues surrounding this particular question were detailed in **Subsection 7.4.1**, where it was highlighted that human and computer marking could not mark responses to *the boy on a swing* version of the question consistently, since the question diagram did not match with the responses.

Q23 asks the test-taker to identify whether a pair of moving blocks ever have the same speed, and to identify when this is if they do. This question was adapted from Q21 of the original FCI, although its wording differs from the original version by referring to the blocks seperately as *"Block A"* and *"Block B"* in the question statement. This change in wording may have caused test-takers to give answers that refer to *"Block A"* and *"Block B"* separately, and these answers were not marked well by the automated marking rules in the Version 1 and Version 3 IRR studies. As a result, the question wording has been reverted to remove references to *"Block A"* and *"Block B"*.

Cases where the question format was changed

Q30 corresponds to Q27 of the FCI, and it asks the test-taker to state what happens to the speed of a moving box after the constant force being applied to it is removed.

The correct answer required contains two parts, *"slows down"* and *"stops"*, and it was difficult to author marking rules that covered both parts of the answer. This difficulty has persisted over several versions of the AMS, so the question was reverted to the multiple-choice version found in the original FCI for the final version of the AMS.

Q33 requires the test-taker to identify the forces acting on a tennis ball after it has been hit by a tennis racquet. It corresponds to Q30 of the FCI, and is notable for being the only question on the FCI which requests test-takers to take *air resistance* into account. Q33 was asked in multiple-response format in Version 1 of the AMS, and in free-response format in both Version 2 and Version 3 of the AMS. However, the free-response versions of this question asked something different than the multiple-choice version of the question that originally appeared on the FCI, as it removes the requirement for *air resistance* to be identified as once of the forces acting on the tennis ball. As a result, Q33 was reverted to a multiple-response version in the final version of the AMS, where it asks the same question as originally intended by the FCI version.

Cases where the question was removed

Two AMS questions were removed for the final version of the AMS because they were consistently found to perform poorly; these questions were Q3 and Q19. Q3 of the AMS asked test-takers to identify the forces acting on a stone after it as been dropped from a building, with the correct answer being *weight* or equivalent. In the case of Q3, almost every test-taker who attempted the question got it right, and this effect occurred across all of the tested cohorts. This raised concerns about what Q3 actually tested, and this implied that Q3 was a poorly designed question. This outcome may have been expected, since Q3 did not appear on the original FCI and was added as an extra question to the AMS, meaning that it had not been rigorously tested previously. As a result, removing Q3 from the AMS was fully justified.

Q19 of the AMS required test-takers to identify the forces acting on an elevator when it was moving up a frictionless shaft. The correct answer required two forces to be identified, *weight* and *tension*, with appropriate synonyms also being acceptable. However, it was found to be difficult to develop consistent marking rules for this question over several versions of the AMS. This was because of the subjectivity involved in deciding what counts as an acceptable synonym for the *tension* force, which translated poorly to the objective computer marking scheme. As was the case for Q3 above, Q19

was added as an extra question to the AMS, and it did not appear in the original FCI. As a result, design oversights may be expected within the question, as it had not previously been thoroughly tested. Removing Q19 from the AMS was hence a justified outcome.

Case where a question was added

Q6 of the final version of the AMS is a multiple-response question that asks students to identify the forces acting on a marble when it is at a point inside a frictionless channel. It corresponds to Q5 of the original FCI, where it was asked as a multiple-choice question. Q6 required students to identify three forces instead of just one or two, and this made developing marking rules for it as a free-response question into an unfeasible exercise, as there were too many variations of the correct and incorrect answers that could be given. In addition, the *rule transfer* strategy used to develop marking rules for free-response questions (previously outlined and discussed in **Chapter 7**) could not be applied to Q6, since there was no similar question on the AMS to inherit the rules from. As a result, Q6 would be used in multiple-response format if it appeared in any version of the AMS.

Q6 was included in Version 1 of the AMS, but was subsequently absent from Version 2 and Version 3. This is because the aim had been to include only free-response questions in Version 2 and Version 3 of the AMS, even though 7 of the questions required students only to type a letter corresponding to their choice of option. As a result, if Q6 had been included in Version 2 and Version 3 of the AMS, it would have been the only multiple-response question amongst the free-response questions. However, the final version of the AMS uses a combination of free-response and selected-response questions, so Q6 was restored to the final version of the AMS. In addition, since Q6 is a question that appeared in the original version of the FCI, restoring it to the final version allows FCI scores to be compared to AMS scores in a consistent way (This is particularly important for further work into the differences between FCI outcomes for different demographic groups). Furthermore, investigating whether it would be possible to develop effective marking rules for Q6 with a significantly larger response set exists as a possible avenue for future work.

8.3.3 Back-testing the final marking rules against the Version 1 responses

To test for consistency, the final AMS marking rules were back-tested against the Version 1 UHM and the Version 2 UHM. The results of the IRR calculations comparing the Version 1 UHM to the final computer marking are given in Table 8.5 below. Note that only 18 questions were in free-response format in the Version 1 AMS, so these are the only questions that have data for the final computer marking to be compared to. In addition, no calculations are presented for Q3, Q11, Q19, Q23 or Q30, as these questions were modified or removed for the final version of the AMS (as explained in **Subsection 8.3.2**).

Question	Marking agreement	Cohen's kappa
Q1	1.00	0.99
Q2	1.00	0.99
Q3	-	-
Q4	0.96	0.91
Q5	0.99	0.98
Q11	-	-
Q13	0.97	0.94
Q17	0.98	0.96
Q19	-	-
Q20	0.99	0.96
Q22	0.99	0.98
Q23	-	-
Q25	0.98	0.96
Q27	0.99	0.96
Q29	1.00	0.99
Q30	-	-
Q31	1.00	0.99
Q32	0.97	0.91

Table 8.5: Table showing the marking agreement and Cohen's kappa values for the Version 1 UHM against the final computer marking rules.

In the results given in Table 8.5, there were no cases where the marking agreement or Cohen's kappa values were outside the acceptable range of values. The Version 1 UHM was not used directly to build the final version of the AMS marking rules (although the overall development of the marking rules was an iterative process), so this finding shows that the final version of the marking rules were highly effective at marking the questions Q1, Q2, Q4, Q5, Q13, Q17, Q20, Q22, Q25, Q27, Q29, Q31

and Q32. This indicated that the automated marking had reached a high level of functionality for these 13 questions.

8.3.4 Back-testing the final marking rules against the Version 2 responses

The results of the IRR calculations which compare the Version 2 UHM to the final computer marking are presented below in Table 8.6. Note that 25 questions were in free-response format in Version 2 of the AMS, so these are the questions that have data for the final computer marking to be compared against. In a similar way to Table 8.5 above, no calculations are given for Q3, Q11, Q19, Q23, Q30 or Q33, as these questions were changed or removed for the final version of the AMS (as previously detailed in **Subsection 8.3.2**).

Question	Marking agreement	Cohen's kappa
Q1	1.00	1.00
Q2	1.00	1.00
Q3	-	-
Q4	0.95*	0.89
Q5	0.99	0.96
Q7	1.00	1.00
Q11	-	-
Q12	0.97	0.94
Q13	0.95	0.86
Q15	0.92*	0.85
Q17	1.00	1.00
Q18	0.95	0.86
Q19	-	-
Q20	0.98	0.96
Q21	-	-
Q22	1.00	1.00
Q23	-	-
Q25	1.00	1.00
Q27	0.98	0.96
Q28	1.00	1.00
Q29	0.97	0.91
Q30	-	-
Q31	1.00	1.00
Q32	0.92*	0.81
Q33	-	-

Table 8.6: Table showing the marking agreement and Cohen's kappa values for the Version 2 UHM against the final computer marking rules.

Based on the calculations presented in Table 8.6, questions Q15 and Q32 were identified as potentially being problematic for the computer marking. In each of these cases, the marking agreement was below the 0.95 cut-off, but the Cohen's kappa was above the 0.8 cut-off, implying that the marking agreement in each of the cases did not arise because of random chance. In addition, there were a smaller number of responses to compare the Version 2 UHM against the final computer marking rules. For example, Q15 was tested against 66 responses, and there were 5 disagreement cases; and Q32 was tested against 60 responses, and 5 disagreement cases occurred. This meant that the effect of having a small number of disagreement cases was inflated when calculating the marking agreement, whereas it did not affect the Cohen's kappa value; this is expected behaviour, since Cohen's kappa is designed to account for such annotator bias (Artstein and Poesio, 2008). As a result, Q4, Q15 and Q32 were not identified to be problematic for the computer to mark.

For the cases of Q1, Q2, Q4, Q5, Q7, Q12, Q13, Q17, Q18, Q20, Q22, Q25, Q27, Q28, Q29 and Q31, the values of both the marking agreement and Cohen's kappa were within the acceptable range of values. Since the Version 2 UHM was not used directly to build the final version of the AMS marking rules (although the overall development of the marking rules was an iterative process), this finding indicates that the final version of the marking rules were functioning at a high level for the 15 free-response questions listed above. In addition, the cases of Q4, Q15 and Q32 were not found to be problematic for the final version of the computer marking. Taking these findings together, 18 out of the 21 free-response questions on the final version of the AMS have marking rules which are consistent enough for general use.

For the questions Q11, Q21 and Q23, issues with the computer marking of the responses were found to be the result of the wording of the questions. Changes were made to the wording of these questions, with the aim of drawing answers from students which the computer is more capable of marking accurately; these changes were explained in detail in **Subsection 8.3.2**. As a result, further testing to verify whether these interventions have had the desired effect on the effectiveness of the computer marking stands as a possible direction for future research.

Summary of the final version of the AMS

The final version of the AMS contains 31 questions. Of these, 21 are free-response; 8 are free-response (letter); and 2 are multiple-response questions. The correspondence between the standardized AMS question numbering; the final AMS version question numbering; the original FCI question numbering; and the question numbering used by the tests on the *Isaac Physics* platform is given in Tables 8.7 and 8.8 below. The final versions of the AMS questions themselves and the corresponding marking rules can be found in **Appendix A**.

Standardized AMS question	Final version AMS question	FCI question	Isaac question	Final version question type
Q1	Q1	Q1	Test 2, Q2, Part 1	Free-response
Q2	Q2	Q2	Test 2, Q2, Part 2	Free-response
Q3	-	Q3	-	-
Q4	Q3	Q3	Test 1, Q2	Free-response
Q5	Q4	Q4	Test 1, Q2	Free-response
Q6	Q5	Q5	Test 1, Q3, Part 1	Multiple-response
Q7	Q6	Q5	Test 1, Q3, Part 2	Free-response
Q8	Q7	Q6	Test 1, Q3, Part 3	Free-response (Letter)
Q9	Q8	Q7	Test 3, Q6	Free-response (Letter)
Q10	Q9	Q8	Test 2, Q4, Part 1	Free-response (Letter)
Q11	Q10	Q9	Test 2, Q4, Part 2	Free-response
Q12	Q11	Q10	Test 2, Q4, Part 3	Free-response
Q13	Q12	Q11	Test 2, Q4, Part 4	Free-response
Q14	Q13	Q12	Test 1, Q4	Free-response (Letter)
Q15	Q14	Q13	Test 2, Q1	Free-response
Q16	Q15	Q14	Test 2, Q3	Free-response (Letter)
Q17	Q16	Q15	Test 2, Q5	Free-response
Q18	Q17	Q16	Test 3, Q4	Free-response
Q19	-	Q17	-	-

Table 8.7: Table showing how the questions on the final version of the AMS map to the standardized AMS question numbering system, the FCI question numbering system, and the Isaac Physics platform question numbering system.

Standardized AMS question	Final version AMS question	FCI question	Isaac question	Final version question type
Q20	Q18	Q17	Test 3, Q3	Free-response
Q21	Q19	Q18	Test 1, Q5	Free-response
Q22	Q20	Q19	Test 3, Q2, Part 1	Free-response
Q23	Q21	Q20	Test 3, Q2, Part 2	Free-response
Q24	Q22	Q21	Test 3, Q5, Part 1	Free-response (Letter)
Q25	Q23	Q22	Test 3, Q5, Part 2	Free-response
Q26	Q24	Q23	Test 3, Q5, Part 3	Free-response (Letter)
Q27	Q25	Q24	Test 3, Q5, Part 4	Free-response
Q28	Q26	Q25	Test 1, Q6, Part 1	Free-response
Q29	Q27	Q26	Test 1, Q6, Part 2	Free-response
Q30	Q28	Q27	Test 1, Q6, Part 3	Free-response (Letter)
Q31	Q29	Q28	Test 1, Q7	Free-response
Q32	Q30	Q29	Test 3, Q1	Free-response
Q33	Q31	Q30	Test 3, Q7	Multiple-response

Table 8.8: Table showing how the questions on the final version of the AMS map to the standardized AMS question numbering system, the FCI question numbering system, and the Isaac Physics platform question numbering system.

8.4 Use of the Isaac Physics platform

There were limitations to the data collected for the IRR testing through the Isaac Physics platform. The main limitation was that some of the questions were tested against a smaller number of responses than in the previous AMS IRR studies (outlined in **Chapter 6** and **Chapter 7**), and this had the effect of amplifying both positive and negative IRR outcomes in the testing. For example, the Version 3 AMS marking rules for Q33 were tested against 32 responses. There were only 2 disagreement cases, and Q33 had a high value for the Cohen's kappa but a slightly lower value for the marking agreement; in this case, a negative effect was amplified despite the small number of disagreements. On the other hand, the Version 3 AMS marking rules for Q22 were tested against 31 responses. There was only 1 disagreement case, and Q22 had high values for both the Cohen's kappa and the marking agreement; this showed a case where a positive effect was amplified by the small number of disagreements. Such cases illustrate the importance of calculating the Cohen's kappa statistic, which is designed to account for the annotator bias (Artstein and Poesio, 2008) which can be responsible for such variations.

The use of the Isaac Physics platform for the administration of Version 3 of the AMS raised further points for discussion. To start, the AMS was split into three tests for use on the Isaac platform, known as *Isaac Test 1*, *Isaac Test 2* and *Isaac Test 3*. Each of the Isaac tests was roughly one third of the length of the AMS, meaning that they could each be completed in a shorter length of the time than the AMS. Unlike the full-length AMS, there did not appear to be any questions on the Isaac tests that had high drop-off rates in participation, so it is possible that having shorter length tests encouraged participants to complete the test after they had started it. However, there are many other factors to consider here, such as differences in the cohorts being tested, so further data would be required to test this claim.

The free-response questions on Isaac Physics made use of the same open source Pattern Match technology as their OSL counterparts, which allowed the marking rules to be directly transferred between the two platforms. All Moodle questions can be easily transferred between different Moodle installations, but the act of moving the AMS to a different external platform demonstrated a wider potential for sharing between platforms. Further work is still required to move the AMS to other external platforms, but the above findings indicate that this is a realistic future direction for

its widespread distribution and use. Whilst there is nothing within the Moodle *Virtual Learning Environment* (**VLE**) per se that limits widespread student access to its tests, a sensitivity around the sharing of FCI questions coupled with the particular functionality of the OSL meant that a potential test-taker needed to create an account in order to attempt the AMS questions on the OSL, which is likely to have reduced the pool of users. The move to Isaac Physics therefore also enabled more widespread access.

The Isaac Physics versions of the AMS questions told students whether the answer that they had entered to the question was marked as correct or incorrect by the computer in real-time, which was not the case on the OSL versions of the questions, although Moodle supports a wide range of feedback options so there was no technical reason for this. Students who did the questions on the Isaac Platform therefore received limited level feedback about their performance, whereas those on the OSL did not. Without further qualitative data it is not possible to know what the Isaac cohort thought about receiving this feedback, or indeed if they found it to be useful. However, the eight students who took the AMS in the usability testing outlined in **Chapter 5** did receive the same basic *correct/incorrect* feedback after they had completed all of the AMS questions, and it was found in the corresponding interviews that they found this feedback to be useful.

The idea of providing automated feedback to free-response questions has previously been explored by Zhu et al. (2020), who found that students in general benefited from receiving instantaneous feedback, regardless of what their level of understanding was. In addition, Zhu et al. used their questions in a formative setting, with the main aim of facilitating learning. In the context of the current work, these findings indicate that free-response questions and educational instruments that make use of them (such as the AMS) could be used as teaching tools instead of being used solely for conceptual evaluation purposes. This idea is further backed by the findings of Bulut et al. (2019), who found that feedback can be used by instructors to help link concepts together and give a holistic view of their courses, which could help students to gauge the level of understanding that is required of them overall. From the above considerations, the idea of combining real-time feedback with the AMS questions could lead to it serving a different purpose, with the focus being oriented more towards enhancing the learning experience of the students.

8.5 Conclusions

The first aim of the study presented here was to develop the AMS marking rules to a high level of functionality, and additional responses to the questions were required to do this. The second aim of the study was to take steps towards expanding the potential user-base of the AMS. Both of these requirements were satisfied by collecting the additional responses to the questions through three shorter-length AMS tests hosted on the *Isaac Physics* platform. The responses were used to build a Version 3 UHM, and this was used to test the performance of the computer marking. As was the case in **Chapter 6** and **Chapter 7**, an *Inter-Rater Reliability* (IRR) approach was used to conduct this testing.

For the Version 3 computer marking, it was found that 20 out of 25 of the free-response questions had acceptable values for both the Cohen's kappa and marking agreement, and 23 out of 25 of the questions had acceptable values for the Cohen's kappa. This implied that the marking rules were operating at a high level of accuracy for most of the free-response AMS questions. In the 2 cases where this was not the case, changes were made to the question wording for the first, and the question was reverted to a multiple-choice question for the second. Considerations from the Version 1, Version 2 and Version 3 AMS studies were used to develop the final version of the AMS, and the final version contained a mixture of selected-response and free-response questions. To check for consistency, the final version marking rules were back-tested against UHMs from previous academic years. It was found that there were no serious problems with the making rules for any of the free-response questions tested on the final version of the AMS, although further testing of questions, and if necessary further development, in particular of questions that have previously been problematic (Q11, Q23 and Q30), is required to ensure that issues do not re-emerge.

Using the Isaac Physics platform allowed the AMS to be attempted by a wider user-base, because it was easier for potential test-takers to access the AMS tests on Isaac Physics than it was on the OSL. The shorter length tests may have encouraged students to complete the test once they had started it, although further data would be required to investigate this idea. The Isaac Physics versions of the AMS questions also gave limited feedback to the test-takers in real-time, by telling them whether their answer had been marked as *correct* or *incorrect* by the computer. As a result of these considerations, the idea of combining the free-response questions of the AMS with

real-time feedback was raised as a possible future direction for the research, as also discussed in **Chapter 5** (especially in **Subsection 5.3.6**).

The findings of **Chapter 6** and **Chapter 7** showed that the AMS questions had a high level of functionality, but the marking rules still required additional work to also reach this same level of functionality. The work conducted here in **Chapter 8** developed the AMS marking rules to a high level of functionality to match with that of the AMS questions, meaning that the AMS as a whole also had a high level of functionality. Going beyond this, the work here expanded the potential user-base of the AMS by moving it to the Isaac Physics platform, and it also highlighted possible alternative uses for the AMS by providing students with a limited amount of real-time feedback on their performance.

8.6 Summary and looking ahead

Chapter 8 presented the quantitative findings from the *Inter-Rater Reliability* (IRR) studies conducted on the Version 3 AMS and final version AMS marking rules, which made use of responses gathered through the *Isaac Physics* platform. It was found that the AMS marking rules were functioning well, and that the AMS can be successfully moved to other platforms. Taken together, these findings illustrate that the AMS has the potential for widespread use.

Chapter 9 outlines the early-stage development of a new concept inventory making use of free-response questions, the *General Relativity Concept Inventory* (GRCI). It presents findings from both qualitative and quantitative data collected from students who attempted the GRCI.

9 The General Relativity Concept Inventory

9.1 Rationale

General Relativity is a mathematical physics subject which has a strong conceptual grounding within Einstein's original thought experiments. Concept inventories have been developed for other mathematical physics subjects, such as Quantum Mechanics (Dick-Perez et al., 2016) and Electromagnetism (Ding et al., 2006; Baily et al., 2017). Further, there exists a Relativity Concept Inventory (RCI) (Aslanides and Savage, 2013) although this tests topics from *Special Relativity*, not *General Relativity*. This leaves a gap within the available resources that could be filled by developing a *General Relativity Concept Inventory* (GRCI). One of the key goals of the overall research is to investigate the use of free-response questions in physics concept inventories, hence it follows that developing a GRCI using free-response questions aligns with the motivations of this research. In addition, it is positively beneficial to explore how free-response concept inventories function for a mathematical physics subject such as General Relativity, as the findings are applicable to a wider Physics Education Research context.

9.2 Methods

Questions for the proposed GRCI needed to be authored in order for the study to be conducted. Material covered in the 1980s OU General Relativity module *S354 Understanding Space and Time* and the modern OU General Relativity module *S383 The Relativistic Universe* was examined to find out which concepts are typically covered in a General Relativity course, as well as to see what types of conceptual and mathematical questions were posed in these modules texts. *S354* was chosen for consideration alongside *S383* because it had been taught from a conceptual standpoint, which the author of this thesis was made aware of through discussions with Professor Robert Lambourne (2017) and Professor Sally Jordan (2017). Based on this material, the following five concepts were chose as the basis for the GRCI:

- The need for a general theory of relativity [CO1].
- The geometry of spacetime [CO2].
- Understanding the Einstein field equations [CO3].
- Predictions and tests of General Relativity [CO4].
- Consequences of General Relativity for Cosmology [CO5].

The first draft of the GRCI contained 20 questions, with four questions corresponding to each of the above listed concepts. Expert review of the questions (Norton, 2018; Serjeant, 2018; Croston, 2018; Lambourne, 2018) deemed that some of the questions were testing *rote knowledge* rather than a deeper conceptual understanding, and it also identified questions which might be difficult to author effective automated marking schemes for. In addition, it was identified that the test was trying to test too many concepts. As a result of this feedback it was decided to reduce the GRCI to 10 questions, testing the following four topics, based on the originally proposed five concepts:

- Frames of Reference [T1].
- Ingredients for General Relativity (Principle of Equivalence and Curvature) [T2].
- The Einstein Field Equations [T3].
- General Relativity and Cosmology [T4].

On this version of the GRCI, there were two questions corresponding to [T1]; three questions corresponding to [T2]; two questions corresponding to [T3]; and three questions corresponding to [T4]. These were the GRCI questions used to conduct the study, and they can be found in **Appendix F**.

9.2.1 Data collection

The objective was to collect meaningful response data to the GRCI questions. Because of the complexity of the subject, the GRCI could only be taken meaningfully by students who were familiar with the concepts of General Relativity in a *post-test* setting. However, since General Relativity is not part of the Core of Physics outlined by the Institute of Physics (IOP, 2020), it is not offered as part of the physics curriculum by all institutions. However, all institutions which offer a General Relativity course should address the same basic concepts covered in the GRCI. Data collection efforts focused on General Relativity courses at three institutions, as these had suitable candidates to attempt the GRCI.

Students from the three different institutions were invited to attempt the GRCI questions. These were students on The Open University's *S383 The Relativistic Universe*; students from the University of Leeds; and students from the University of Cambridge. All the students used in the study were close to the end of an undergraduate degree, on a physics pathway. This meant that the students were all of similar

levels of study experience, and had the same subject backgrounds. However, the participants would have had different levels of prior exposure to the topics in the GRCI, based on what topics had been covered on each particular course. No conscious effort was made to gather students of different ethnicities or genders, since investigating specific demographics was not the aim of this study, and this would have been impractical, given the small number of participants. In addition, feedback was not given to the GRCI test-takers about their performance; this was because investigating giving feedback on concept inventories was not an aim of the current study, unlike the AMS study conducted previously in **Chapter 5**.

The study made use of 26 participants in total. Testing and interviewing a greater number of participants would have been useful, but there were not large numbers of students available to do this. This is acknowledged here as a limitation of the study. However, the methods chosen to analyze the data are considered (Braun and Clarke, 2006) to be reliable for small numbers of participants, so the numbers in this study were deemed to be sufficient.

Because of different requirements and availability of the cohorts, each of the institutions administered the GRCI in a different way, and these are outlined here. For the students from The Open University, the GRCI questions were administered online through the Pattern Match question type on the OSL platform, in much the same way as the different versions of the AMS were. The Pattern Match GRCI questions had only basic marking rules, derived from the author's experience and understanding of the literature. Nine students completed the GRCI in this way, and the data were collected by downloading them directly from the OSL. For the students from the University of Leeds, the GRCI questions were completed on paper under exam conditions, during a lecture; 13 students completed the GRCI in this way, and the scripts were posted to The Open University where the results were manually collected together into a spreadsheet. For the students from the University of Cambridge, the GRCI questions were handed to the participants to do on paper, and a short interview about the experience followed; four students completed the GRCI in this way, and they each received an *Amazon* voucher worth £20 in appreciation of their involvement. The written GRCI responses were manually collected together into a spreadsheet, and the interview responses were transcribed manually. For reference, the differences in administration methods at each of the institutions are summarized in Table 9.1 below. The

interview questions asked to the four participants from the University of Cambridge can be found in **Appendix H**.

Institution	Number of participants	Was the GRCI paper based or computer based?	Was the GRCI taken under exam conditions?	Were the participants interviewed about their experience?
OU	9	Computer based	No	No
University of Leeds	13	Paper based	Yes	No
University of Cambridge	4	Paper based	Yes	Yes

Table 9.1: Table showing the different methods used to administer the GRCI at the various institutions involved in the study.

There were two components to the data. The first were the written answers given by the 26 participants to the GRCI questions; the second were the verbal responses given by the four participants from the University of Cambridge to the interview questions. These two types of data formed a rich data set, and both parts needed to be treated in a different way. The GRCI interview responses were treated as qualitative data, with the objective of finding out what the participants thought of the GRCI. The written responses were treated as quantitative data, and were used to find out about the concepts that students used when giving correct and incorrect answers; this information was used to develop early stage marking rules for the GRCI questions. Note that 26 sets of responses were not sufficient for more advanced development of the marking rules to take place, so this served only as a *proof of principle* investigation into the authoring of effective marking rules for the GRCI questions.

9.2.2 Data analysis

The GRCI response data from the 26 participants formed one data set. The responses were put into a spreadsheet, and marked against the model answers by the author. As was the case for the AMS responses, the marking was binary, with a mark of 1 awarded for an answer that was judged to be correct, and a mark of 0 awarded for an answer that was judged to be incorrect. In line with practice described earlier in this thesis, the *correct* answers were examined to see if any positive marking rules could be developed, whereas the *incorrect* answers were examined to see if any negative

marking rules could be developed. The objective was to use the small number of responses gathered to iterate the basic marking rules for the GRCI questions that the author had previously written. Further, the responses themselves were examined to see what sort of misconceptions the participants demonstrated when answering the questions.

The GRCI interview data formed another data set, and *Thematic Analysis* (Braun and Clarke, 2006; Braun et al., 2014) was used to find underlying themes. As in **Chapter 5**, Thematic Analysis can be used to analyze a qualitative data set drawn from a small number of participants being involved in a study. Thematic Analysis reduces this data into a form which can be interpreted, since the data are difficult to understand in their raw form; this results in the eponymous *themes* of the method. The approach prevents the investigator from drawing arbitrary conclusions from the data, since themes are distilled from the rate of occurrence of their underlying codes within the data set. The same version of Thematic Analysis used in **Chapter 5** (University of Auckland, 2017) was used to analyze the data here.

9.3 Results and Discussion

9.3.1 Findings from the GRCI responses

The common key concepts used in correct answers to the GRCI questions and the common key misconceptions used in incorrect answers to the GRCI questions are considered below. Using these answers, the same approach employed to develop the marking rules for the AMS (discussed previously in **Subsection 6.2.2** and **Subsection 6.4.2**) was used to iterate the author's rudimentary set of marking rules for each GRCI question, and these updated rules can be found in **Appendix F**. In addition, the questions from the GRCI referred to in the following discussion can also be found in **Appendix F**.

The topic of *Frames of Reference* was tested in Q1 and Q2 of the GRCI. For Q1, correct answers often referred to Newtonian gravitation being an approximation, for example *"That it must be an approximation and that it will break down when relativistic effects are important"*. Incorrect answers often repeated information from the question without reaching the required conclusion about Newtonian gravitation, such as *"Newtonian gravitation does not take the same form in all inertial frames of reference"*. For Q2, correct answers referred to the observer being in a non-accelerating

state of motion, for example *"The observer is not accelerating. That they moving at constant velocity, or at rest"*. Incorrect answers were unable to reach the conclusion about the observer's state of motion.

The topic of *Ingredients for General Relativity (Principle of Equivalence and Curvature)* was tested in Q3, Q4 and Q5 of the GRCI. For Q3, correct answers commonly referred to the observer being unable to tell whether their motion was due to the effects of gravity or accelerating motion, whereas incorrect answers arose because students were unable to properly interpret the Principle of Equivalence, for example *"The observer cannot tell that they are accelerating"*. For Q4, correct answers frequently referred to the Curvature of the non-Euclidean space; incorrect answers instead named a specific kind of geometry without giving the general result, such as spherical geometry. For Q5, correct answers identified that Curvature gives rise to the non-Euclidean spaces required for the geometric view of spacetime used in General Relativity, whereas incorrect answers such as *"Curves"* did not manage to make this realization.

The topic of *The Einstein Field Equations* was tested in Q6 and Q7 of the GRCI. For Q6, correct answers noted that the Einstein Field Equations determine the geometry of the spacetime; whereas incorrect answers failed to recognize this, such as *"It relates to gravitational potential"*. For Q7, correct answers recognized that the energy-momentum tensor determines the matter-energy inventory of the spacetime; incorrect answers were instead unable to identify this, for example by relating the energy to the momentum, but not to the spacetime.

The topic of *General Relativity and Cosmology* was tested in Q8, Q9 and Q10 of the GRCI. For Q8, correct answers were able to identify that applying the Cosmological Principle to make the homogeneous and isotropic assumptions simplifies the situation, for example *"Homogeneous and isotropic is a simplification that is adopted to prevent the equations becoming impossible to calculate"*. Incorrect answers were not able to capture this idea, such as *"Because it is necessary that the same laws of physics apply in all regions of the universe"*. For Q9, correct answers noted that the matter-energy density of the Universe is thought to be at the critical density; whereas incorrect answers, such as *"Most energy Is dark energy"* did not recognize this. For Q10, correct answers noted that it is currently thought that the Universe will expand forever; on the other hand, incorrect answers typically listed various different possible scenarios for the fate of the Universe, such as *"Heat Death or Big Crunch (Depending on k)"*.

Of note, *"Heat Death"* alone would have been an acceptable answer, but adding the *"or"* aspect invalidated this answer, since the student was presenting more than one scenario.

The marking rules that were authored based on the correct and incorrect answers to each of the questions were back-tested against the human-marked responses used to author them. The calculated marking agreement for each of the GRCI questions is shown in Table 9.2 below.

GRCI question	Topic tested	Marking agreement
Q1	Frames of Reference	0.85
Q2	Frames of Reference	0.77
Q3	Ingredients for General Relativity	0.92
Q4	Ingredients for General Relativity	0.92
Q5	Ingredients for General Relativity	0.92
Q6	The Einstein Field Equations	0.92
Q7	The Einstein Field Equations	0.88
Q8	General Relativity and Cosmology	0.92
Q9	General Relativity and Cosmology	0.96
Q10	General Relativity and Cosmology	0.92

Table 9.2: Table showing the marking agreement values of the GRCI marking rules against the GRCI human marker.

From the results in Table 9.2, Q2 had a particularly low back-testing marking agreement. This occurred because there were several different responses to Q2, and these did not group together in a consistent way to develop marking rules from. As a result, this means that the question might need re-wording to scaffold students towards giving answers which are suitable to be automatically marked. This is in line with the previously discussed idea from the literature that computer marking can be made more effective by changing question wording, as well as by modifying the automated marking schemes (Butcher and Jordan, 2010).

Aside from the case of Q2 discussed above, the marking agreement values in Table 9.2 are all 0.85 or above, which in an encouraging result for the development of automated mark schemes for the free-response questions of the GRCI. However, at least several hundred further responses would be required to further develop and test these marking rules (Jordan and Mitchell, 2009). In addition, the results in Table 9.2 were obtained through back-testing, meaning that it is likely that they have been over-fitted (Zehner et al., 2016). The GRCI marking rules were not further developed in this study; as a next step, the rules would need to be tested against a new set of responses.

Discussion of GRCI questions and responses

Various points for discussion were raised by the development and pilot use of the GRCI. When initially authoring the questions for the GRCI, 20 questions were proposed, although only 10 of these managed to make it into the version of the GRCI used throughout the study. In line with standard Delphi process used to develop concept inventories (Porter et al., 2014), some questions were removed because when reviewed by experts they were found only to test rote knowledge. Further, some proposed GRCI questions were not used because they could not be put into a conceptual form as a result of of their mathematical grounding, which is an issue that has previously arisen in the literature with the development of the BEMA concept inventory for Electromagnetism (Ding et al, 2006).

The questions on the GRCI were based on the course content from the OU Relativity modules *S354 Understanding Space and Time* and *S383 The Relativistic Universe*, which meant that the GRCI questions aligned better with OU courses than those at other Universities, and this was a clear limitation of the study. However, there are fundamental concepts that are likely to be included in all General Relativity curricula, and the questions on the GRCI were authored based around this design priority. The GRCI questions were found to be general enough for participants from other institutions to answer them in the intended way when used in the current study. However, if institutions offer more advanced General Relativity courses (such as Cosmology), then it would be possible for them to expand the GRCI (so add more questions to it) for use in these courses. The idea of developing concept inventories to match curriculum requirements was previously taken up by Baily et al. (2017) in their development of the CURrENT concept inventory for Electrodynamics, which illustrated that the

approach is viable.

The written responses to the GRCI questions revealed common understandings and misunderstandings of the students. The incorrect answers to Q1 and Q2 revealed that students struggled to properly understand Frames of Reference. This topic has previously been identified as being important for the overall understanding of General Relativity in the work of Semon et al. (2009), because it provides a historical and conceptual link between Special Relativity and General Relativity. The incorrect answers to Q3 revealed that students were unable to apply the Principle of Equivalence to solve a specific problem, and this was consistent with the previous findings of Bandyopadhyay and Kumar (2010).

Q4 and Q5 were based around Curvature, whereas Q6 and Q7 were based on the Einstein Field Equations; hence each of these questions tested General Relativity topics that were mathematical in nature. Q4, Q5, Q6 and Q7 were generally well-answered by the students, which may be expected because General Relativity courses are often mathematically grounded (Hartle, 2008). The incorrect answers to Q8, Q9 and Q10 showed that students had difficulties interpreting the results of General Relativity in the Cosmology context. This result agrees with the work of Conlon et al. (2017), where it was found that students have various ideas about the fate of the Universe.

The findings from the written responses to the GRCI questions can be summarized as follows. Incorrect answers to Q1, Q2 and Q3 showed that students found the physical interpretation of General Relativity to be difficult; correct answers to Q4, Q5, Q6 and Q7 showed that students had a good understanding of the mathematical aspects of General Relativity; and incorrect answers to Q8, Q9 and Q10 showed that students struggled to apply the results of General Relativity to the Cosmology context. The students answering the mathematical type questions better than the physical type questions may be a reflection of the General Relativity instruction that students had received, and this consideration could be useful for the development of concept inventories for other mathematical physics subjects.

Although the process used to develop the GRCI questions was largely successful, possible limitations with the approach were also found. To start, there were instances in the GRCI questions (Q8 and Q10) where students were scaffolded towards giving a particular answer to the scenario. In the case of Q8, the question does not make a direct

reference to the *Cosmological Principle* in the reasons for making the *homogenous* and *isotropic* assumptions in the formulation of Cosmological models; whereas in the case of Q10, one particular Cosmological model was selected to base the correct answer upon. This restricts the free-response aspect of the questions, and should be considered in any possible further revisions to the questions.

Further, General Relativity is a broad and advanced subject area, which means that selecting key concepts to base the GRCI questions on is a subjective process. This is because experts in different areas of the field (such as theoretical physicists and observational cosmologists) may select different concepts to be the *key concepts* of General Relativity, which could lead to them developing very different versions of the GRCI based upon their own biases; such arguments also apply to other mathematical physics subjects such as Electromagnetism and Quantum Mechanics. As a result, developing concept inventories for certain subjects may be more difficult because of their broad content, as well as because of their inherently mathematical nature.

Twenty-six sets of responses were used to develop *proof of principle* computer marking rules. All the questions except Q2 had sufficiently consistent correct responses to develop plausible positive marking rules, but only Q9 and Q10 had sufficiently consistent incorrect responses to develop negative marking rules. It should however be noted that negative marking rules are not necessarily required for automated marking to be effective; most incorrect responses are handled simply by not being matched against a positive marking rule. However, the marking rules were developed and tested using only these 26 sets of responses, which meant that the marking rules could easily become *over-fitted* to match the limited data set (Zehner et al., 2016). Hundreds more responses to the GRCI questions would be required to further develop and test the marking rules (Jordan and Mitchell, 2009), and this is a possible avenue for future development of the GRCI.

9.3.2 Findings from the GRCI interviews

The *Thematic Analysis* conducted on the GRCI interview data identified 7 codes and 1 sub-code, and these grouped together into 3 themes: *Free-response questions are appropriate to test for conceptual understanding of General Relativity*, *General Relativity joins physical interpretations with mathematical constructs, and both are important when fluent with the theory*, and *The GRCI can be given a formative purpose.* In what follows, each of these themes are discussed, with supporting quotes from the interviews referred to as required. In addition, the four participants interviewed are labelled as P1, P2, P3 and P4 for the remainder of this chapter.

Findings from the *Free-response questions are appropriate to test for conceptual understanding of General Relativity* theme

There were 2 codes associated with the *Free-response questions are appropriate to test for conceptual understanding of General Relativity* theme, and this theme was coded 26 times overall. The codes associated with this theme are presented in Table 9.3 below. Unless otherwise stated, the occurrence of the codes was more or less equal between the four participants. Furthermore, each code is assigned a label (such as C1) such that they can easily be referred to in the text. These conventions are used in Tables 9.3, 9.4 and 9.7.

Code	Number of times coded
Free-response questions made them think (C1)	21
Free-response questions test understanding more thoroughly than multiple-choice questions (C2)	5

Table 9.3: Codes associated with the *Free-response questions are appropriate to test for conceptual understanding of General Relativity* theme.

For code C1, the participants noted that the open-ended format of the free-response GRCI questions was different from the General Relativity questions that they were accustomed to. In addition, they noted that the free-response question type made them think about their answers, since a *semantic* approach was required to come up with the words to answer with:

"I think it's good to get, like, have this open-ended way, trying to formulate in your own words. Then, when I'm trying to write down these sentences, I can really feel myself being tested, like, do I understand it? Because I have to come up with it" [P1].

In code C2, the participants contrasted the free-response format of the GRCI questions with that of multiple-choice questions. They noted that free-response questions prevent test-takers from making use of *eliminate and guess* strategies to reach the correct answer:

"With multiple-choice, it would be, look at the answers, and find the one that fits best" [P3].

It is of note that the participants appeared to be aware that questions were typically asked in multiple-choice format, which might suggest that they had some previous exposure to concept inventories, or were at least aware of them.

Through codes C1 and C2, participants identified that the free-response format was appropriate for testing for conceptual understanding of General Relativity, since the open-ended format gave them the chance to express what they were really thinking. Participants felt that having to come up with their own answers was better suited for testing their General Relativity understanding than multiple-choice questions. This was because for the free-response questions, the participants identified that they had to think carefully about what they knew, and write their own words to articulate these ideas; this was contrasted with the multiple-choice format, where they would have instead selected from a list of options that somebody else had already written.

In line with the above considerations, one participant went into further detail about how they approached different question types more generally. They reflected that mathematical questions can be solved in a procedural manner; whereas multiple-choice questions can be answered using an *eliminate and guess* strategy; and free-response questions require students to tap into their knowledge of the subject to be answered. This was an example of the participant being a *conscientious consumer* (Higgins et al., 2002), as they reacted to questions in the way that they believed they were supposed to. The above reflections were similar to those given by the participants from AMS usability laboratory study reported in **Chapter 5**, and this indicates that free-response questions can be used to test for conceptual understanding of various physics topics.

Findings from the *General Relativity joins physical interpretations with mathematical constructs, and both are important when fluent with the theory* theme

There were 4 codes related to the *General Relativity joins physical interpretations with mathematical constructs, and both are important when fluent with the theory* theme, and these are given below in Table 9.4. Overall, this theme was coded 40 times. An additional sub-code (labelled as SC1) was referred to exclusively by participant P1, but it was retained for the analysis because it raised an interesting point about the appeal of General Relativity as a subject.

Code	Number of times coded
In responding to the GRCI questions, participants referred to the mathematical interpretations of General Relativity (C3)	5
In responding to the GRCI questions, participants referred to the physical interpretations of General Relativity (C4)	29
In responding to the GRCI questions, there were cases where the answer was difficult to articulate using words (C5)	6
It was recognized that General Relativity could be used in the context of science engagement (SC1)	2

Table 9.4: Codes associated with the *General Relativity joins physical interpretations with mathematical constructs, and both are important when fluent with the theory* theme.

In code C3, the participants made reference to the mathematical interpretation of General Relativity, although they noted that no calculations were required to answer the GRCI questions:

"You're already so mathematically versed that you literally might think in terms of equations, when actually, you might want to try and understand it proper" [P4].

For code C4, participants identified that they needed to understand what they were writing about in their answers, since they were being made to think about the physical interpretations of the theory. In addition, participants identified that it was important to understand how the theory was originally formulated, as well its physical and philosophical implications:

"That was the main thing that I got from that, actually trying to think about the physical reasoning" [P1].

Related to this, participants identified that the physical interpretations of General Relativity could be difficult to put into words, which was captured through code C5:

"I was also expecting it to be tricky, because I find - I think a lot of us find trying to explain what you're doing is quite hard, in words" [P1].

Through codes C3 and C4, participants noted that General Relativity is a mathematical and abstract topic, although it also has physical consequences that are applicable in the everyday world. In code C5, participants found that it was difficult to express these mathematical and physical ideas in words. Subsequently, sub-code SC1 related codes C3, C4 and C5 to a *science engagement* context, where one participant noted that General Relativity would be difficult to explain to a non-specialized audience. This was because the audience would not have the fluency in mathematics required to understand General Relativity, but also because the subject is difficult to explain in words:

"It's all well and good being able to do this kind of algebra, but what does that mean for the lay person? How does that really affect things in the real world? Why does that mean we see things as they are? That's almost more important...and it's difficult, it really is difficult" [P1].

Through Code C3, participants noted that the GRCI questions were not mathematical, which contrasted with their previous experience of answering General Relativity assessment questions. Code C4 highlighted that the questions on the GRCI got the participants to look beyond the mathematics of the situations, and to think about what was going on from a physical standpoint. This was identified as being a challenging task in code C5, but was also seen as rewarding because it helped the participants to see the physical consequences of the theory, and to give it some context in a *bigger picture* view. Related to this point, participants further noted that being tested on General Relativity conceptual understanding forced them to go back and think about how the theory was formulated, what the key concepts really meant, and how they fitted together to form the coherent theory of General Relativity. These reflections showed that some of the participants thought of General Relativity mainly in mathematical terms, which highlighted their weaknesses when it came to interpreting and

understanding the results in a physical context. Taken together, these points indicated that doing the GRCI was a useful exercise for the participants.

Bloom's Taxonomy can be used to understand the above points pertaining to gathering and building General Relativity knowledge. Bloom's taxonomy is a classification system for educational objectives, and was originally proposed as a one-dimensional construct (Bloom, 1956). The taxonomy was revised (Krathwohl, 2002) to include the two dimensions of *knowledge* and *cognitive process*, and the various tiers of these dimensions are shown in Tables 9.5 and 9.6 below.

Within the context of Bloom's revised Taxonomy, the importance of conceptual understanding for the mastery of General Relativity is highlighted by the *Conceptual knowledge* level from Table 9.5 and the *Understand* level from Table 9.6. However, conceptual understanding of General Relativity is underpinned by having mathematical competence with the techniques of the subject, as well as being able to interpret results in a physical setting. Referring again to the revised Taxonomy, the mathematical competences required to solve General Relativity problems are underpinned by the *Procedural knowledge* level from Table 9.5 and the *Apply* level from Table 9.6; whereas the capacity to interpret the results of such calculations in a physical context pertains to the *Analyze* level from Table 9.6. Students with less expertise may struggle with one or both of the mathematical and physical aspects of the subject, which hinders their conceptual understanding of General Relativity. However, Bloom's revised taxonomy is a *continuum*, which means that such students have the potential to improve their conceptual understanding by studying the subject and mastering the mathematical and physical aspects; this mastery comes when the student gains the ability to move between both interpretations, and can explain these to others.

Knowledge level	Meaning
Factual knowledge	The student can recall basic facts related to the subject.
Conceptual knowledge	The student is able to put different elements of the subject into a knowledge structure.
Procedural knowledge	The student can apply methods to solve problems within the subject.
Metacognitive knowledge	The student is self-aware of their own levels of knowledge and cognition.

Table 9.5: Table showing the different levels of the *knowledge* dimension of Bloom's revised Taxonomy.

Cognitive process level	Meaning
Remember	The student can recall basic elements related to the subject.
Understand	Student is able to determine meaning from recalled elements related to the subject.
Apply	The student can carry out a method or procedure to solve a problem related to the subject.
Analyze	The student can see how different parts of the subject fit together as part of a *bigger picture* view.
Evaluate	The student makes judgements based on their own understanding of the subject matter.
Create	The student puts the different pieces of the subject together to form a coherent worldview.

Table 9.6: Table showing the different levels of the *cognitive process* dimension of Bloom's revised Taxonomy.

In the current GRCI study, the participants noted that General Relativity is a highly mathematical subject, and that they had been taught it in a manner that reflected this. However, participants also contrasted the mathematical aspects of the subject with the way that General Relativity can be applied physically in the everyday world, which illustrated that the participants were aware that both mathematical and physical interpretations were important for understanding of the theory. This line of thought led one participant to muse about using General Relativity in scientific engagement ventures, which was captured within sub-code SC1. This was an important point, because General Relativity is a key part of understanding space and time, but it is not well understood by the majority of people. Indeed, General Relativity is often not taught as part of the undergraduate physics curriculum (Hartle, 2008), meaning that only those who specialize in the subject are ever likely to gain any exposure to it. It follows that applying the GRCI, or a tool similar to it, to a science engagement context is a possible novel direction for future work.

Findings from the *The GRCI can be given a formative purpose* theme

The *The GRCI can be given a formative purpose* themes was coded 34 times overall. The theme consists of 2 codes, and these are shown in Table 9.7 below.

Code	Number of times coded
Participants used the experience of doing the GRCI to reflect upon their understanding of General Relativity (C6)	29
Within an educational context, participants identified that the GRCI could be used as a teaching tool (C7)	5

Table 9.7: Codes associated with the *The GRCI can be given a formative purpose* theme.

For code C6, the participants each talked through the answers they gave to the GRCI questions in some detail. Each participant reflected upon how well they had done, and noted that they had not answered questions well in some instances. In some cases, participants admitted that they had forgotten the content required to answer the questions, since their corresponding exams on the topic of General Relativity had already occurred. In addition, all of the participants identified that they would have liked to get feedback on their work:

"I guess that it's good to know kind of, where I am in my understanding of General Relativity. I don't know if I said it like nonsense...so it would be nice to know...I feel like feedback would be most useful in here" [P2].

Further, the participants discussed how the GRCI could be used for teaching purposes in a General Relativity course, which was captured in code C7:

"There would actually be room for discussion in a supervision [tutorial] regarding use of questions...if you say, 'right, this is, like a, no-pressure sort of test, we just want to see where your understanding is...don't look it up...just think about it off the top of your head, and it can be useful for us to focus in on what bits you conceptually understand, and which bits you don't really, which need more focus'. Yeah, that's what I'd say" [P3].

Code C6 illustrated that all of the participants reflected on how well they had answered the GRCI questions. Some of the participants felt as if they had forgotten the

General Relativity content required to answer the questions because they had studied the topics several months previously, and in some instances the participants were not confident in their answers as a result of this. In addition, participants typically talked through their line of reasoning on questions that they had found difficult. All of the participants felt as if it would have been useful to get feedback on the GRCI questions. This was because the participants wanted to know whether their answers were correct, and to highlight gaps in their understanding. This finding is similar to that from the AMS usability laboratory study covered in **Chapter 5**, and it agrees with the idea from the literature that students generally like getting feedback on their work (Kluger and DeNisi, 1996; Brown and Glover, 2006; Zhu et al., 2020).

Looking ahead, some of the participants went on to discuss possible future uses for the GRCI, and these instances were captured within code C7. One participant suggested using the GRCI in a small group setting to find weaknesses and build conceptual knowledge; this would allow the GRCI to be used as a teaching tool, which is a different function to that of a standard concept inventory. This notion was previously explored in **Chapter 7** with respect to using the AMS as a teaching tool with feedback, and this links back to the idea from the literature that feedback could be used to facilitate with conceptual understanding (Bulut et al. 2019).

Several attempts to teach General Relativity within different contexts have been detailed in the literature. For example, Burko (2017) proposed a General Relativity teaching activity which involved students analyzing gravitational wave data from LIGO; Muller and Frauendiener (2011) developed an interactive tool for studying the geodesics of various spacetimes; and Zahn and Kraus (2014) designed a workshop to teach about curved spacetimes using models of black holes. On a more conceptual level, Kaur et al. (2017) developed an analogies-based approach for teaching General Relativity concepts at the school level. In particular, Kaur et al. believed that it was important for Einsteinian physics to be taught in schools, because of its real-world applications and its importance to modern physics. These examples indicate that there exist contexts where the GRCI could be deployed as a teaching tool, and further development and testing of the GRCI towards this objective stands an avenue for further work.

9.4 Conclusions

The aim of the study was to investigate the use of free-response questions to test conceptual understanding of General Relativity topics. To this end, 10 General Relativity conceptual questions were authored, and were put together into a *General Relativity Concept Inventory* (GRCI). Data were gathered by giving the GRCI to 26 participants to complete, and by interviewing four of the participants about their experience of doing the GRCI. For the automated marking aspect of the study, it was found that basic marking rules could be developed for the 10 GRCI questions using the 26 sets of responses. Further development and testing of these rules is required, but the study provided a *proof of principle* for the idea of developing other concept inventories using free-response questions.

The written responses revealed that participants struggled with the questions based on the physical topics of Frames of Reference and the Principle of Equivalence; they did well on the questions based on the mathematical topics of Curvature and Einstein's field equations; and they struggled with the questions which required them to interpret General Relativity results in a Cosmology setting. These findings were consistent with the observation that most General Relativity courses cover the subject from a mathematical standpoint. This highlighted the potential for the GRCI to be used as a teaching tool in General Relativity courses, encouraging the subject to be viewed from a deeper conceptual standpoint.

Although largely successful, limitations were also identified with the process used to develop the GRCI. First, it was noted that some of the GRCI questions were posed in such a way as to scaffold test-takers towards giving particular answers. Second, it was postulated that it may be difficult to develop concept inventories for General Relativity and other mathematical physics subjects, since they are very broad in their coverage. Taking these considerations into account could help the guide the development of concept inventories for other mathematical physics subjects in the future.

The interview responses showed that participants felt that free-response questions were suitable for testing conceptual understanding of General Relativity. The participants further noted that the questions required no mathematics to answer, and they contrasted this with the mathematical General Relativity questions with which they were familiar. This led participants on to discuss the nature of General Relativity as an abstract and difficult subject, and the idea of using General Relativity in science

engagement ventures was also raised. Finally, participants reflected upon their performance on the GRCI questions, and they thought that getting feedback from the GRCI would facilitate with their learning of the subject. Taken together, these findings indicate that the GRCI is a useful educational tool that encourages students to use concepts, and it could be developed to further support the learning of General Relativity in a teaching context. More generally, this use stands as a possible future direction for use of free-response questions in physics concept inventories.

9.5 Summary and looking ahead

Chapter 9 presented the findings from the GRCI study. The GRCI responses allowed a basic set of marking rules to be developed for the questions, and this served as *proof of principle* for the idea of computer marking for the GRCI questions. The written responses to the GRCI questions showed that participants struggled with questions that required physical interpretations of the subject, but did well on questions which covered mathematical topics. The GRCI interview responses found that free-response questions were suitable for testing conceptual understanding of General Relativity, and participants noted that General Relativity has both mathematical and physical interpretations. In addition, participants thought that getting feedback from the GRCI would facilitate with their learning of the subject, and the idea of using the GRCI in the teaching context was discussed.

Chapter 10 summarizes the overall findings from the research, revisits the research questions, and looks at possible future directions for the research.

10 Conclusions and future work

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.
T.S. Eliot, *Little Gidding*

The aim of this final chapter is to summarize the main conclusions from the previous chapters, and to use these to answer the original research questions posed in **Section 1.2**. It finishes by looking ahead to possible avenues for future research that have been opened up by the work, as well as giving a reflection on the future of concept inventories as educational instruments in an increasingly technology-driven world.

10.1 Summary of the research findings

The research carried out in the previous chapters is summarized below. Taken together, these highlight key design priorities for physics concept inventories that make use of free-response questions.

Chapter 1 introduced the project, and outlined the research questions that would be the focus of the work in the thesis. **Chapter 2** reviewed the literature related to concept inventories and automated marking of free-response questions. It was detailed that concept inventories in physics are instruments commonly used to assess the conceptual understanding of students. In addition, most of these concept inventories make use of multiple-choice questions in their construction, which limits the amount of information that can be obtained about students' genuine level of understanding of physics concepts. To counter this effect, it was proposed that the questions could instead be asked in the free-response format, which would allow students to write their own answers. However, to efficiently make use of this question format, the responses would need to be automatically marked because of the marking burden incurred by the written answers. These ideas underpinned the research carried out in the rest of the thesis.

Chapter 3 presented a case study of the multiple-choice FCI being used in practice. The findings were that students improved their performance on the FCI after studying the relevant Newtonian mechanics material, but they often still struggled with the

concept of Newton's Third Law. The FCI (which makes use of multiple-choice questions) was adapted into the AMS (which makes use of free-response questions), and the research outcomes from developing and testing the AMS formed the main part of the studies carried out in the remainder of the thesis. As an introduction to this, **Chapter 4** outlined the resources and process used to develop and test the AMS, as well as how the questions on the different versions of the AMS mapped to one another.

Chapter 5 presented the qualitative findings from the usability study conducted with Version 1 of the AMS, as well as the findings from the responses given to the qualitative feedback question (Q34). It was found that participants could see the educational value of answering free-response questions instead of multiple-choice, and they responded well to these questions. In addition, students welcomed feedback on their performance after working through the AMS questions, and most were seen to make use of it to help them learn.

Chapter 6 presented the quantitative findings from the Classical Test Theory (CTT) and Inter-Rater Reliability (IRR) studies conducted using responses gathered to Version 1 of the AMS. The findings indicated that the AMS questions were functioning well, but the marking rules still required further development. **Chapter 7** continued the development and testing of the AMS, and presented findings from the Classical Test Theory (CTT) and Inter-Rater Reliability (IRR) studies conducted using responses gathered to Version 2 of the AMS; this version of the AMS contained more free-response questions than the previous version. The findings from the CTT strand of the study indicated that the AMS questions were functioning well, whereas the findings from the IRR strand of the study showed that the corresponding AMS marking rules still required further development. The student cohorts in Version 1 and Version 2 were identified as being different, with the latter group having had, on average, less previous exposure to the ideas of Newtonian mechanics. Since concept inventories are likely to attract a wide range of users, this was considered an important part of the development process.

Chapter 8 concluded the development process of the AMS, and presented the findings from the Inter-Rater Reliability (IRR) studies conducted on the Version 3 AMS and final version AMS marking rules, which made use of responses gathered through the Isaac Physics platform, another characteristically different group of users. It was found that the AMS marking rules were now functioning well, and that the

AMS could be successfully moved from the OSL to other hosting platforms. Although some further testing would be sensible as the user base grows, to check for unforeseen unmatched responses, the findings of the current work point towards a strong potential for widespread use of the AMS as an alternative to the multiple-choice FCI.

Chapter 9 expanded the approach used to construct the AMS to build the GRCI, a concept inventory assembled using free-response questions based on the subject of General Relativity. Responses to the GRCI were used to develop a basic set of computer marking rules for the questions, although these still required further development and testing. The written responses to the GRCI showed that participants answered questions which covered mathematical topics well, but struggled with questions which required physical interpretations of the theory. Responses in the GRCI interviews found that free-response questions were suitable for testing conceptual understanding of General Relativity, and participants noted that the subject has both mathematical and physical aspects. Furthermore, participants identified that getting feedback from the GRCI would facilitate with their learning, and the notion of using the GRCI for teaching purposes was raised.

This section has summarized the research conducted in the thesis, and it is now important to go back and see how this can be used to answer the corresponding research questions. This is the focus of the next section.

10.2 Answering the research questions

The research questions outlined in **Section 1.2** were as follows:

• RQ1: To what extent are free-response versions of a physics concept inventory questions valid and reliable?

• RQ2: How reliable are automated marking schemes when used to mark free-response concept inventory questions?

• RQ3: How effective are concept inventories when used to assess the conceptual understanding of a mathematically involved physics subject?

In what follows, each of the research questions is answered in the context of the work completed in the thesis. At this point, it is worth recalling that the approach used to answer **RQ1** and **RQ2** made use of the FCI, as this is the most well-known physics

concept inventory and has previously been rigorously tested. Since the approach was found to work for the FCI, it could well be applied to other physics concept inventories too. Similarly, the approach used to answer **RQ3** used the subject of General Relativity because there was a gap in the available resources, and it could be applied to develop concept inventories for other mathematical physics subjects also.

RQ1: To what extent are free-response versions of a physics concept inventory questions valid and reliable?

RQ1 was answered by authoring free-response versions of FCI questions, and putting these together into a new concept inventory known as the AMS. The AMS is different from the traditional multiple-choice FCI, so the AMS needed to be tested for validity, to see whether it was capable of doing what it was designed to do. This validity testing was carried out through the usability testing and corresponding interviews presented in **Chapter 5**. It was found that free-response versions of the FCI questions could be used in place of the multiple-choice questions, making the approach valid. Because the questions on the AMS are of a different format than those on the FCI, the AMS questions also needed to be tested for reliability, to check whether they were consistent. This reliability testing was carried out through the Classical Test Theory (CTT) analysis presented in **Chapter 6** and **Chapter 7**. This analysis found that the AMS questions generally functioned well, making the AMS questions reliable. Taking these results together, the AMS questions were found to be valid and reliable.

The research findings related to **RQ1** can be summarized as follows. Qualitative testing of the AMS revealed that students could see the educational value of using free-response questions instead of multiple-choice, and students responded well to answering the free-response questions. As a result, the free-response AMS questions could be used in place of the corresponding multiple-choice FCI questions, which validated the approach. In addition, students welcomed receiving feedback after working through the AMS, and made use of it to reflect upon their performance on it. Quantitative testing of the AMS questions revealed that the different cohorts tested had different levels of performance on the AMS, and that there were different trends and behaviours observed in questions which tested similar concepts. In addition, issues were identified with the distractors on some of the multiple-choice questions, as some of these distractors were rarely selected by students. Furthermore, the quantitative testing found that the AMS questions were reliable; taking this together with the above qualitative finding that

the AMS questions were valid, this showed that the AMS questions were functioning well. In turn, these research findings raised further questions (labelled below as **FQs**) about the use of the FCI and the AMS, and these are listed below:

• **FQ1:** Do different groups of students (such as different demographic groups) respond differently to free-response and multiple-choice questions on physics concept inventories?

• **FQ2:** Can the detailed responses to the AMS free-response questions be used to resolve queries about the distractor options in the FCI multiple-choice questions?

• **FQ3:** Are there situations where it is more appropriate to use selected-response questions to test for conceptual understanding?

• **FQ4:** How might concept inventories serve an expanded purpose in the future?

RQ2: How reliable are automated marking schemes when used to mark free-response concept inventory questions?

RQ2 was answered by authoring marking rules for the free-response versions of the FCI questions used in the AMS, and this was done through the Pattern Match question type of the Moodle question engine. These marking rules needed to be tested for reliability, to see if they were capable of accurate marking that was consistent with that of expert human markers. This reliability testing was carried out in the Inter-Rater Reliability (IRR) analysis presented in **Chapter 6**, **Chapter 7** and **Chapter 8**, and it was found that the AMS marking rules generally functioned well, meaning that the AMS marking rules were reliable.

The research findings related to **RQ2** can be summarized as follows. Quantitative testing of the AMS marking rules highlighted some cases where AMS free-response questions were difficult for the computer to mark accurately, as well as other cases whether the AMS free-response questions were difficult for the human markers to mark consistently. In addition, there were cases where human and computer markers would be expected to outperform one another, owing to the properties of the questions and the responses that were given to them. Furthermore, the quantitative testing found that the AMS marking rules were reliable, which illustrated that the AMS marking rules were functioning well. The research findings raised further questions (**FQs**) about the use of automated marking in concept inventories, and these are listed below:

• **FQ5:** Does scaffolding questions to make them suitable for automated marking turn these questions into selected-response questions?

• **FQ6:** Is it preferable to have more false positives or more false negatives when using automatic marking of free-response questions in concept inventories?

• **FQ7:** Can the rule authoring process outlined in this study be effectively automated?

• **FQ4:** How might concept inventories serve an expanded purpose in the future? (This **FQ** was also raised by the findings of **RQ1** above)

RQ3: How effective are concept inventories when used to assess the conceptual understanding of a mathematically involved physics subject?

RQ3 was answered by authoring questions designed to test for conceptual understanding of General Relativity topics. These questions were put together into a draft version of the General Relativity Concept Inventory (GRCI), and the work carried out to develop and test the GRCI was presented in **Chapter 9**. It was found that the GRCI was a useful educational tool which encouraged students think about the concepts, and that the GRCI could potentially be used within a teaching setting in the future.

The research findings related to **RQ3** can be summarized as follows. Analysis of the GRCI responses showed that students did well on the questions based on mathematical aspects, but struggled with the questions based on interpreting the theory in a physical context. Qualitative testing of the GRCI revealed that students thought that free-response questions were suitable for testing conceptual understanding of General Relativity, and they also identified that the mathematical and physical interpretations of General Relativity were both important in order to understand the subject. In addition, students reflected upon their performance on the GRCI questions, and students felt that getting feedback would facilitate with their learning of the subject. These findings raised further questions (**FQs**) about using concept inventories to assess the conceptual understanding of mathematically involved physics subjects, and these are listed below:

• **FQ8:** With a larger response set, is it possible to develop a version of the GRCI which is both reliable and valid?

• **FQ9:** Is the approach used to develop the GRCI effective when used with other mathematical physics subjects such as Quantum Mechanics and Electromagnetism?

• **FQ10:** When testing conceptual understanding of mathematical physics subjects such as General Relativity, is it possible to fully separate the mathematics from the physics?

• **FQ4:** How might concept inventories serve an expanded purpose in the future? (This **FQ** was also raised by the findings of **RQ1** and **RQ2** above)

This section has revisited the research questions and shown how they were answered, and the research findings from each of them have highlighted other possible areas for future work. The final section looks ahead to see how some of this work may be carried out, as well as reflecting upon the possible future use of concept inventories.

10.3 Possible future work

The FCI and the AMS were designed to measure students' conceptual understanding of Newtonian mechanics. However, the written responses to the AMS questions brought up the idea that the FCI and AMS could be capable of measuring other constructs as well. This point was previously broached in the early days of the FCI in the work of Huffman and Heller (1995), who claimed that the FCI tested mastery of different force-related situations, but was not a test of the force concept itself. It is generally agreed within the PER community that the FCI is capable of measuring something, but there are still disagreements about what this something may be (Wallace and Bailey, 2010). This has in turn led to a variety of studies based around the FCI. For example, Scott and Schumayer (2017) have used the FCI to learn about the conceptual coherence of students' non-Newtonian worldviews; whereas Ishimoto et al. (2017) compared the FCI performances of Japanese and American students; and the work outlined in this thesis investigated the use of free-response versions of the FCI questions.

Martin-Blas et al. (2010) found that students with different levels of previous exposure to physics reacted differently to the FCI questions. This phenomenon was also observed in students who took the AMS, as the different cohorts all had different previous exposures to physics, and performed differently on the questions. The idea of investigating differences in AMS performance by demographic group was not explored in this thesis, but it raises a question (**FQ1**) which could be a possible future direction

for the research. In particular, whether the free-response format of the AMS questions can be used alongside a scaffolding approach (Dawkins et al., 2017) to reduce the gender **attainment gap** in physics would be a viable start point for future work.

A strength of multiple-choice questions is that they can be efficiently and accurately marked by both computer and human markers. However, this comes at the cost of forcing students to pick from a list of pre-prepared options, and these may not reflect the lines of thought employed by the students. In the research presented, there were examples of free-response AMS questions where students used other lines of reasoning to the FCI distractors, and other cases where some distractors were rarely selected on AMS multiple-choice questions. This raised the question (**FQ2**) of whether the distractors in some of the FCI questions were the *most appropriate* distractors. In the literature, this idea has previously been investigated by Rebello and Zollman (2004) and Yasuda et al. (2018), although these studies focused on the distractors of a few FCI questions only. As a result, a future line of research could focus on completing a typology of all of the distractors on the FCI by comparing them to typed free-response answer data gathered using the AMS.

The research also raised a question (**FQ3**) about situations where automatically-marked free-response questions were not suitable to test for student understanding. Jordan and Mitchell (2009) noted that questions required specific qualities to be suitable for marking rules to be authored for them. Such questions need to assess objective constructs, and there needs to be a small number of feasible correct answers that can be given to them. Through careful consideration and development, all of the free-response questions on the final version of the AMS fitted this criterion. Conversely, Jordan and Mitchell found it was difficult to author marking rules for questions where there were too many different ways to give the correct answer, and for cases where a lot of the responses contained both correct and incorrect content. This final point was relevant to one AMS question, which was eventually reverted to multiple-choice format after several unsuccessful attempts to author marking rules to disentangle the correct and incorrect answers. In addition, AMS questions which involved the identification of a trajectory were found to be better suited to the selected-response format, since it can be difficult to accurately describe a trajectory using words.

Two of the free-response AMS questions were scaffolded in a way that encouraged students to give certain answers, and this raised another question (**FQ5**) about

whether such questions were effectively selected-response in format. A similar point was raised by Sarrouti and El Alaoui (2020) who postulated that the accuracy of automated marking would be improved by setting up questions to return exact answers, which could be achieved by having students answer questions which required a specific word; a *yes/no* response; or *true/false* response. However, this approach was not suitable for the work carried out in this thesis, because the objective was to learn about students' understanding and misunderstandings by giving them the freedom to express themselves using their own words. Indeed, the effectiveness of the automated marking is directly affected by the wording of the question being asked (Butcher and Jordan, 2010), so any attempts to drastically improve the marking agreement in troublesome cases should start with an attempt to re-word the question itself. The above considerations and design priorities are relevant to any future work which makes use of automatically-marked free-response questions to test for conceptual understanding of physics.

Since computer marking can never be flawless, there will inevitably be cases where *false negatives* and *false positives* arise in the automated marking of free-response questions. Another question raised by the research (**FQ6**) was whether it was preferable to have more false positives or false negatives when using automatically-marked, free-response concept inventories. False negatives mean that students who understand a topic do not get the credit they deserve for answering the question. As a result, summative assessment should aim to avoid false negatives, as these could potentially lower a student's overall course grade. Conversely, false positives mean that students who do not understand a topic get credit for answering the question, and this can give students a false sense of their own level of understanding. For the work carried out in this thesis, reducing the number of false negatives was given priority over reducing the number of false positives, but the opposite approach could also have reasonably been applied instead. Prioritizing the reduction of false positives or false negatives appears to be a decision that is taken based upon context, and is an important consideration to make when making use of automatically-marked free-response questions.

The work presented in the thesis made improvements towards the automatic marking of free-response questions. The rule creation process employed was a manual process whereby human-marked and computer-marked responses were compared to find instances of false positives and false negatives and these cases were used to develop corresponding marking rules. The process proved to be effective, although it was time-

consuming. This raised the question (**FQ7**) of whether the rule creation process could be effectively automated. Automating the process would save time, and it would also make use of the Pattern Match question type more accessible for those who do not have expertise in the required syntax and logic, which would allow the technology to be more widely used. The **AMATI** feature of the Pattern Match question type (Willis, 2010) was a step towards automating the rule creation process, but it remains as an avenue for future work. For example, a follow-up study to the work presented in this thesis could focus on comparing the effectiveness of automatically generated computer marking schemes to human written computer marking schemes for the AMS questions, and the results could highlight strengths and weaknesses in the two types of marking scheme.

Free-response questions were also used in the development of the GRCI in this thesis. The questions on the GRCI were written newly for the instrument, and only 26 sets of responses went into developing the computer marking rules for the GRCI. It is a natural follow-up question (**FQ8**) to ask whether the GRCI could be developed into a valid and reliable instrument with a larger response set. To test the GRCI questions for validity, *usability testing* similar to that detailed in **Chapter 5** could be conducted; whereas to test the GRCI questions for reliability, the Classical Test Theory (CTT) statistics calculated in **Chapter 6** and **Chapter 7** could be used. As a result, conducting validity and reliability testing on the GRCI questions stands as a possible direction for future work. To develop and test the GRCI marking rules, the Inter-Rater Reliability (IRR) approach used in **Chapter 6**, **Chapter 7** and **Chapter 8** could be employed. For the size of the response set, the previous work of Jordan and Mitchell (2009) suggested that several hundred responses are typically required to develop accurate marking rules for each free-response Pattern Match question. However, the GRCI questions are new, meaning that extra responses may be required to check if there is anything wrong with the questions to start with. The pilot study into developing automated marking schemes for the GRCI questions (covered in **Chapter 9**) indicated that the idea was feasible, although more responses would be required to investigate this feasibility in practice. Developing the automatic marking of the GRCI questions to be more reliable is hence another possible direction for future work.

General Relativity is a mathematical physics subject, and a question raised by the research (**FQ9**) was whether the approach used to develop the free-response questions of the GRCI could be used to develop concept inventories for other mathematical

physics subjects such as Quantum Mechanics and Electromagnetism. Previous efforts to develop concept inventories for Electromagnetism include the BEMA (Ding et al., 2006) and the CURrENT (Baily et al., 2017). However, the BEMA asks mainly mathematical questions with extra distractor options rather than conceptual Electromagnetism questions. Further, Baily et al. (2017) recognized that testing conceptual understanding of Electromagnetism is difficult because of the mathematics involved. In the case of Quantum Mechanics, Dick-Perez et al. (2016) developed the QCCI to test for conceptual understanding of Quantum Chemistry topics. Notably, Dick-Perez et al. believed that it was possible to measure conceptual understanding for mathematical subjects such as Quantum Mechanics. The work in this thesis to develop the GRCI echoes this view, although the conceptual grounding of General Relativity within Einstein's original thought experiments may have made the task of authoring conceptual questions for General Relativity more straightforward than it would be for Quantum Mechanics or Electromagnetism.

A related question raised by the research (**FQ10**) was whether it was possible to fully separate the mathematics from the physics in a subject such as General Relativity. General Relativity is typically taught in a mathematical setting (Hartle, 2008), but there have been several attempts within the General Relativity Education literature to teach General Relativity within different physical contexts (Burko, 2017; Muller and Frauendiener 2011; Zahn and Kraus, 2014) and from a conceptual standpoint (Kaur et al., 2017). However, the different levels of *knowledge* and *cognitive process* outlined in Bloom's revised Taxonomy (Krathwohl, 2002) indicate that the separation of General Relativity into physical interpretations and mathematical competences may come with higher levels of expertise; this is reflected in the idea that experts of the subject (such as cosmologists) are required to connect concepts together, interpret situations from a physical standpoint, and solve problems using mathematical methods. As a comparison, students who have just started to learn the subject will have less expertise, and will start at the lower levels of the Taxonomy; these students can advance to the higher levels by familiarizing themselves with both the physical interpretations of General Relativity and the procedural mathematics required to solve problems within the subject. Using the GRCI to investigate differences in thinking between students with various levels of experience hence stands as a possible avenue for further work.

Concept inventories are traditionally used to test the effectiveness of teaching methods by administering them as a pre-test and a post-test (Bailey et al., 2012). One final

over-arching question (**FQ4**) raised by the research is how concept inventories may be used in the future. The research has developed physics concept inventories that make use of free-response questions which are automatically marked, and these can be used in different ways to traditional concept inventories. Because of their online delivery, it is possible for the questions to give feedback in real-time to students once the computer marking is sufficiently accurate to enable feedback. Expanding this feature would open up the option of using concept inventories as teaching tools, and would also facilitate students to manage and regulate their own learning (Boud and Soler, 2016). Developing and investigating the inclusion of automatically-generated feedback in free-response physics concept inventories is another potential avenue for further investigation.

Pattern Match proved to be very successful resource for the work carried out in this thesis. However, a limitation of the Pattern Match approach to automatically marking free-response questions is that it cannot be directly translated into other languages. For instance, in languages such as Slovak (Bocková, 2019) and Tagalog (Parker, 2019; Pineda, 2019) the way in which meaning is extracted from sentences is different from English, meaning that the Pattern Match approach of searching for particular words, features and structures cannot be used to differentiate between correct and incorrect answers in these languages. Hence for other languages, different approaches would be required. In spite of this, the techniques of pattern recognition have successfully been applied in the wider linguistics context for the purpose of translating between different languages (Papineni et al., 2001; Doddington, 2002; Qin et at., 2009). Beyond this, pattern recognition has a wide range of other applications such as passport control, search engines, drug recognition (Liu et al., 2015), and galaxy classification (Banerji et al., 2010; Gauci et al., 2010; Kuminski et al., 2014; Schutter and Shamir, 2015) which makes the findings of the current research highly relevant in an increasingly technology-dominated landscape.

Returning to the FCI, the research presented in this thesis has established the AMS, a physics concept inventory that makes use of free-response questions which are marked automatically, and has shown that the AMS is a valid and reliable instrument. As a result, the AMS provides a viable alternative to the paper-based, multiple-choice version of the FCI. The world is no longer the same as it was when the FCI was first used in 1992; there are new technologies available, new assessment techniques have been developed, and the students of today have a different worldview to their

predecessors. Concept inventories need to adapt for use in an ever-changing world, and it has now been shown that it is achievable to move them to an automatically marked, post-multiple-choice era. As a result, the concept inventories of tomorrow could well be vastly different to those of yesterday.

11 References

Alfonseca, E. and Perez, D. (2004) 'Automatic Assessment of Short Questions with a Bleu-inspired Algorithm and shallow NLP', *Lecture Notes in Computer Science*, vol. 3230, pp. 25-35.

Ali, S. H., Carr, P. A. and Ruit, K. G. (2016) 'Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter', *Journal of the Scholarship of Teaching and Learning*, vol. 16, no. 1, pp. 1-14.

Artstein, R. and Poesio, M. (2008) 'Inter-Coder Agreement for Computational Linguistics', *Computational Linguistics*, vol. 34, no. 4, pp. 555-596.

Aslanides, J. S. and Savage, C. M. (2013) 'Relativity Concept Inventory: Development, Analysis, and Results', *Physical Review Special Topics Physics Education Research*, vol. 9, no. 1.

Bacon, D. R. (2003) 'Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marketing context', *Journal of Marketing Education*, vol. 25, no. 1, pp. 31-36.

Bailey, J. M., Johnson, B., Prather, E. E. and Slater, T. F. (2012) 'Development and Validation of the Star Properties Concept Inventory', *International Journal of Science Education*, vol. 34, no. 14, pp. 2257-2286.

Baily, C., Ryan, Q. X., Astolfi, C. and Pollock, S. J. (2017) 'Conceptual assessment tool for advanced undergraduate electrodynamics', *Physical Review Physics Education Research*, vol. 13.

Bandyopadhyay, A. and Kumar, A. (2010) 'Probing students' understanding of some conceptual themes in general relativity', *Physical Review Special Topics Physics Education Research*, vol. 6.

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., Thomas, D. and Vandenberg, J. (2010) 'Galaxy Zoo: reproducing galaxy morphologies via machine learning', *Monthly Notices of the Royal Astronomical Society*, vol. 406, no. 1, pp. 342-353.

Barnum, C. B. (2010) *Usability testing essentials: ready, set, test*, Morgan Kaufmann Publishers, Burlington, Massachusetts.

Bates, S., Donnelly, R., MacPhee, C., Sands, D., Birch, M. and Walet, R. (2013) 'Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison', *European Journal of Physics*, vol. 34.

Ben-Shakter, G. and Sinai, Y. (1991) 'Gender differences in multiple-choice tests: The role of different guessing tendencies', *Journal of Educational Measurement*, vol. 28, no. 1., pp. 23-25.

Betts, L. R. and Elder, T. J., Hartley, J. and Truman, M. (2009) 'Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes?', *Assessment and Evaluation in Higher Education*, vol. 34, no. 1, pp. 1-15.

Bloom, B. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*, New York, McKay.

Bocková, J. (2019). Personal communication.

Boud, D. and Soler, R. (2016) 'Sustainable assessment revisited', *Assessment and Evaluation in Higher Education*, vol. 41, no. 3, pp. 400-413.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77-101.

Braun, V., Clarke, V. and Terry, G. (2014) 'Thematic analysis', *Qualitative Research in Clinical Health Psychology*, vol. 24, pp. 95-114.

Bridgeman, B. (1991) 'A comparison of open-ended and multiple-choice question formats for the quantitiatve section of the Graduate Record Examinations General Test', *ETS Research Report Series*, vol. 2, pp. 1-25.

Brown, D. E. (1989) 'Students' concept of force: the importance of understanding Newton's third law', *Physics Education*, vol. 24, no. 6.

Brown, E. and Glover, C. (2006) 'Evaluating written feedback', in Bryan, C and Clegg, K. (eds) *Innovative Assessment in Higher Education*, London, Routledge, pp. 81-91.

Bull, J. and McKenna, C. (2000) 'Quality assurance of computer-added assessment: Practical and strategic issues', *Quality assurance in Education*, vol. 8, no. 1, pp. 24-31.

Bull, J. and McKenna, C. (2004) *Blueprint for computer-aided assessment*, London, RoutledgeFalmer.

Bulut, O., Cutumisu, M., Aquilina, A. M. and Singh, D. (2019) 'Effect of Digital Score Reporting and Feedback on Students' Learning in Higher Education.', *Frontiers in Education*, vol. 4.

Burko, L. M. (2017) 'Gravitational Wave Detection in the Introductory Lab', *The Physics Teacher*, vol. 55, pp. 288-292.

Burton (2005) 'Multiple-choice and true/false tests: myths and misapprehensions', *Assessment and Evaluation in Higher Education*, vol. 30, no. 1, pp. 65-72.

Butcher, P. (2008) 'Online assessment at the Open University using open source software: Moodle, OpenMark and more', *12th International CAA Conference*, Loughborough, United Kingdom.

Butcher, P. and Jordan, S. (2010) 'A comparison of human and computer marking of short free-text student responses', *Computers and Education*, vol. 55, pp. 489-499.

Carless, D. (2006) 'Differing perceptions in the feedback process', *Studies in Higher Education*, vol. 31, no. 2, pp. 219-233.

Castro, A. and Andrews, G. (2018) 'Nursing lives in the blogosphere: A thematic analysis of anonymous online nursing narratives', *Journal of Advanced Nursing*, vol. 74, no. 2, pp. 329-338.

Chen, J. C., Kadlowec, J., Whittinghill, D. (2004) 'Work in progress: combining concept inventories with rapid feedback to enhance learning', *34th Annual Frontiers in Education*.

Clement, J. (1982) 'Students' Preconceptions in Introductory Mechanics', *American Journal of Physics*, vol. 50, no. 1, pp. 66-71.

Clow, D. (2013) 'An overview of learning analytics', *Teaching in Higher Education*, vol. 18, no. 6, pp. 683-695.

Cohen, J. (1960) 'A coefficient for nominal scales', *Educational and Psychological Measurement*, vol. 20, pp. 37-46.

Conlon, M., Coble, K., Bailey, J. M. and Cominsky, L. R. (2017) 'Investigating undergraduate students' ideas about the fate of the Universe', *Physical Review Physics Education Research*, vol. 13.

Conole, G. and Warburton, B. (2005) 'A review of computer-assisted assessment', *Research in Learning Technology*, vol. 13, no. 1, pp. 17-31.

Crisp, G. (2007) *The e-assessment Handbook*, London, Continuum.

Crocker, L. and Algina, J. (1986) *Introduction to classical and modern test theory*, California, Wadsworth Group/Thompson Learning.

Cronbach, L. J. (1951) 'Coefficient Alpha and the Internal Structure of Tests', *Psychometrika*, vol. 16, pp. 297-334.

Croston, J. (2017). Personal communication.

Croston, J. (2018). Personal communication.

Cuff, B., Robertson, D. and Keys, E. (2019) 'Automated Scoring of Essays and Short Answers - A review of Common Methods', *Ofqual report.*

D'Avanzo, C. (2008) 'Biology Concept Inventories: Overview, Status, and Next Steps', *BioScience*, vol. 58, no.11, pp. 1079-1085.

Draaijer, S., Jordan, S. and Ogden, H. (2018) 'Calculating the random guess scores of multiple-response and matching test items', *Proceedings of the 2017 International Technology Enhanced Assessment Conference (TEA 2017)*, Barcelona, Spain.

Dawkins, H. R., Hedgeland, H. and Jordan, S. (2017) 'Impact of scaffolding and question structure on the gender gap', *Physical Review Physics Education Research*, vol. 13.

Dedic, H., Rosenfield, S. and Lasry, N. (2010) 'Are all wrong FCI answers equivalent?', *AIP conference proceedings.*

Dick-Perez, M., Luxford, C. J., Windus, T. L. and Holme, T. (2016) 'A Quantum Chemistry Concept Inventory for Physical Chemistry Classes', *Journal of Chemical Education*, vol. 93, no. 4, pp. 605-612.

Ding, L., Chaby, R., Sherwood, B. and Beichner, R. (2006) 'Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment', *Physical Review Special Topics - Physics Education Research*, vol. 2.

Ding, L. and Beichner, R. (2009) 'Approaches to data analysis of multiple-choice questions', *Physical Review Special Topics Physics Education Research*, vol. 5.

Ding, L. and Caballero, M. D. (2014) 'Uncovering the hidden meaning of cross-curriculum comparison results on the Force Concept Inventory', *Physical Review Special Topics Physics Education Research*, vol. 10, no. 2.

Docktor, J. and Heller, K. (2008) 'Gender differences in both Force Concept Inventory and Introductory Physics Performance', *AIP Conference Proceedings*.

Docktor, J. L. and Mestre, J. P. (2014) 'Synthesis of discipline-based education research in physics', *Physical Review Special Topics-Physics Education Research*, vol. 10, no. 2.

Doddington, G. (2002) 'Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics', *Human Language Technology Conference*, San Diego, California, United States of America.

Doran, R. (1980) *Basic Measurement and Evaluation of Science Instruction*, NSTA, Washington DC, United States.

Downing, S. M. (2003) 'Guessing on selected-response examinations', *Medical Education*, vol. 37, no. 8, pp. 670-671.

Draper, S. (2009) 'What are learners actually regulating when given feedback?', *British Journal of Educational Technology*, vol. 40, no. 2, pp. 306-315.

Dufresne, R. J., Leonard, W. J. and Gerace, W. J. (2002) 'Marking sense of students answers to multiple-choice questions', *The Physics Teacher*, vol. 40, pp. 174-180.

Eaton, P., Willoughby, S. D. (2018) 'Confirmatory factor analysis applied to the Force Concept Inventory', *Physical Review Physics Education Research*, vol. 14.

Eaton, P., Vavruska, K. and Wiloughby, S. (2019) 'Exploring the preinstruction and postinstruction non-Newtonian world views as measured by the Force Concept Inventory', *Physical Review Physics Education Research*, vol. 15.

Eden, S. (2018). Personal communication.

Epstein, J. (2013) 'The Calculus Concept Inventory - Measurement of the Effect of Teaching Methodology in Mathematics', *Notices of the American Mathematical Society*, vol. 60, no. 8, pp. 1018-1026.

Ferrao, M. (2010) 'E-assessment within the Bologna paradigm: evidence from Portugal', *Assessment and Evaluation in Higher Education*, vol. 35, no. 7, pp. 819-830.

Filia, K. M., Jackson, H. J., Cotton, S. M., Gardner, A., Killackey, E. J. and Cook, J. A. (2018) 'What is Social Inclusion? A Thematic Analysis of Professional Opinion', *Psychiatric Rehabilitation Journal*, vol. 41, no. 3, pp. 183-195.

Funk, S. C. and Dickson, K. L. (2011) 'Multiple-choice and short-answer exam performance in a college classroom', *Teaching of Psychology*, vol. 38, no. 4, pp. 273-277.

Galloway, R. (2019). Personal communication.

Garcia-Quiroga, M. and Hamilton-Giachritsis, C. (2017) 'Getting involved: A thematic analysis of caregivers' perspectives in Chilean residential children's homes', *Journal of Social and Personal Relationships*, vol. 34, no. 3, pp. 356-375.

Gauci, A., Adami, K, Z. and Abela, J. (2010) 'Machine Learning for Galaxy Morphology Classification', *Monthly Notices of the Royal Astronomical Society.*

Gipps, C. and Murphy, P. (1994) *A fair test? Assessment, achievement, and equality*, Open University Press, Buckingham.

Gray, K., Rebello, N. S. and Zollman, D. (2002) 'The effect of Question Order on Responses to Multiple-choice Questions', *2002 Physics Education Research Conference.*

Grogan, S. and Jayne, M. (2017) 'Body image after mastectomy: A thematic analysis of younger women's written accounts', *Journal of Health Psychology*, vol. 22, no. 11, pp. 1480-1490.

Gwinnett, C. and Cassella, J. (2011) 'The trials and tribulations of designing and utilising MCQs in HE and for assessing forensic practitioner competency', *New Directions in the Teaching of Physical Sciences*, vol. 7, pp. 72-78.

Hake, R. R. (1998) 'Interactive-engagement vs traditional methods: a six-thousand student survey of mechanics test data for introductory physics courses', *American Journal of Physics*, vol. 66, pp. 64-74.

Halloun, I. and Hestenes, D. (1985) 'The initial knowledge state of college students', *American Journal of Physics*, vol. 53, pp. 1043-1056.

Halloun, I., Hake, R., Mosca, E. and Hestenes, D. (1995) 'Force Concept Inventory (revised 1995)', in Mazur, E. (1997) *Peer Instruction: A Users Manual*, Prentice-Hall, New Jersey.

Han, J., Bao, L., Chen, L., Cai, T., Pi, Y., Zhou, S., Tu, Y. and Koenig, K. (2015) 'Dividing the force concept inventory into two equivalent half-length tests', *Physical Review Special Topics-Physics Education Research*, vol. 11, no. 1.

Han, J., Koenig, K., Cui, L., Fritchman, J., Li, D., Sun, W., Fu, Z. and Bao, L. (2016) 'Experimental validation of the half-length Force Concept Inventory', *Physical Review Special Topics Physics Education Research*, vol. 12, no. 2.

Hartle, J. B. (2008) 'General Relativity in the Undergraduate Physics Curriculum', *American Journal of Physics*, vol. 74, no. 1, pp. 14-21.

Henderson, R., Miller, P., Stewart, J., Traxler, A. and Lindell, R. (2018) 'Item-Level Gender Fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism', *Physical Review Physics Education Research*, vol. 14.

Henderson, R., Stewart, J. and Traxler, A. (2019) 'Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism', *Physical Review Physics Education Research*, vol. 15.

Hestenes, D., Wells, M. and Swackhamer, G. (1992) 'Force concept inventory', *The Physics Teacher*, vol. 30, pp. 141-158.

Hestenes, D. and Halloun, I. (1995) 'Interpreting the Force Concept Inventory: A Response to March 1995 Critique by Huffman and Heller', *The Physics Teacher*, vol. 33.

Higgins, R., Hartley, P. and Skelton, A. (2002) 'The conscientious consumer: Reconsidering the role of assessment feedback in student learning', *Studies in Higher Education*, vol. 27, no. 1, pp. 53-64.

Hudson, R. D. (2010) 'Multiple-choice questions compared to short-answer responses: Which assess understanding of chemistry more effectively?', *PhD dissertation*, Curtin University, Australia.

Huffman, D. and Heller, P. (1995) 'What does the force concept inventory actually measure?', *The Physics Teacher*, vol. 33, pp. 138-143.

Hughes, M. J. (2002) 'How I misunderstood Newtons third law', *The Physics Teacher*, vol. 40.

Hunt, T and Jordan, S. (2016) 'I wish I could believe you: the frustrating unreliability of some assessment research', *Practitioner Research in Higher Education*, vol. 10, no. 1, pp. 13-21.

IOP website (2020). Available at https://www.iop.org/education/higher_education/accred-and-recog/file_73757.pdf (Accessed 27th July 2020).

Isaac Physics website (2020). Available at https://isaacphysics.org/ (Accessed 27th July 2020).

Ishimoto, M., Davenport, G. and Wittmann, M. (2017) 'Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students', *Physical Review Physics Education Research*, vol. 13.

Itza-Ortiz, S. F., Rebello, N. S. and Zollman, D. (2003) 'Students' Models of Newtons Second Law in Mechanics and Electromagnetism', *European Journal of Physics*, vol. 25.

James, R. L. (2017). Personal communication.

Johnstone, A. H. and Mughol, A. R. (1976) 'Concepts of Physics at Secondary Level', *Physics Education.*

Jordan, S. (2009) *Investigating the use of short free-text questions in online assessment, COLMSCT Final Report*, The Open University, Milton Keynes, United Kingdom.

Jordan, S. and Mitchell, T. (2009) 'e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback', *British Journal of Educational Technology*, vol. 40, no. 2, pp. 371-385.

Jordan, S. (2012a) 'Student engagement with assessment and feedback: some lessons from short-answer free-text e-assessment questions', *Computers and Education*, vol. 58, no. 2, 818-834.

Jordan, S. (2012b) 'Short-answer e-assessment questions: five years on', *Proceedings of the 2012 International Computer Assisted Assessment (CAA) Conference*, Southampton, United Kingdom.

Jordan, S. (2013) 'E-assessment: past, present and future', *New Directions in the Teaching of Physical Sciences*, vol. 9, no. 1, pp. 87-106.

Jordan, S. (2017). Personal communication.

Jorion, N., Gane, B. D., James, K., Schroeder, L., Dibello, L. V. and Pellerino, J. W. (2015) 'An Analytic Framework for Evaluating the Validity of Concept Inventory Claims', *Journal of Engineering Education*, vol. 104, no. 4, pp. 454-496.

Kaur, T., Blair, D., Moschilla, J., Stannard, W. and Zadnik, M. (2017) 'Teaching Einsteinian Physics at Schools: Part 1, Models and Analogies for Relativity'.

Klein, R., Kryrilov, A. and Tokman, M. (2011) 'Automated assessment of short free-text responses in computer science using latent semantic analysis', *Proceedings of the 16th annual joint conference on innovation and technology in computer science education*, pp. 158-162.

Kline, P. (1986) *A Handbook of Test Construction: Introduction to Psychometric Design*, Methuen, London.

Kluger, A. N. and DeNisi, A. (1996) 'The effects of feedback interventions on performance: A historical view, a meta-analysis and a preliminary feedback intervention theory', *Psychological Bulletin*, vol. 119, no. 2, pp. 254-284.

Krathwohl, D. R. (2002) 'A Revision of Blooms Taxonomy: An Overview', *Theory Into Practice*, vol. 41, no. 4.

Kuminski, E., George, J., Wallin, J. and Shamir, L. (2014) 'Combining Human and Machine Learning for Morphological Analysis of Galaxy Images', *Publications of the Astronomical Society of the Pacific*, vol. 126, no. 944, pp. 959-967.

Kuder, G. F. and Richardson, M. W. (1937) 'The theory of the estimation of test reliability', *Psychometrika*, vol. 2, no. 3, pp. 151-160.

Lambourne, R. (2017). Personal communication.

Lambourne, R. (2018). Personal communication.

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A. and Reshef, O. (2011) 'The puzzling reliability of the Force Concept Inventory', *American Journal of Physics*, vol. 79, no. 9, pp. 909-914.

Laverty, J. T. and Caballero, M. D. (2018) 'Analysis of the most common concept inventories in physics: What are we assessing?', *Physical Review Physics Education Research*, vol. 14.

Lawrie, G., Wright, A., Schultz, M., Dargaville, T., OBrien, G., Bedford, S., Williams, M., Tasker, R., Dickson, H. and Thompson, C. (2013) 'Using formative feedback to identify and support first-year chemistry students with missing or misconceptions. A practice report', *International Journal of the First Year in Higher Education*, vol. 4, no. 2, pp. 111-116.

Leacock, C. and Choddorow, M. (2003) 'C-rater: Automated Scoring of Short-Answer Questions', *Computers and Humanities*, vol. 37, no. 4, pp. 389-405.

Lindell, R. S., Peak, E. and Foster, T. M. (2007) 'Are they all created equal? A comparison of different concept inventory development methodologies', *AIP Conference Proceedings*, pp. 14-17.

Liu, S., Tang, B., Chen, Q. and Wang, X. (2015) 'Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. Manually constructed dictionaries', *Information (Switzerland)*, vol. 6, no. 4, pp. 848-865.

Mackintosh, R. (2017). Personal communication.

Madsen, A., McKagen, S. B. and Sayre, E. C. (2013) 'Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?', *Physical Review Special Topics: Physics Education Research*, vol. 9.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S. and Rizvi, M. (2017) 'Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions', *Annals of PIMS*, vol. 13, no. 4.

Martin-Blas, T., Seidel, L. and Serrano-Fernandez, A. (2010) 'Enhancing Force Concept Inventory diagnostics to identify dominant misconceptions in first-year engineering physics', *European Journal of Engineering Education*, vol. 35, no. 6, pp. 597-606.

Martinez, B. and Perez, J. (2010) 'ITEMS Project: An online sequence for teaching mathematics and astronomy', *AIP Conference Proceedings*, vol. 1283, no. 1, pp. 161-165.

Martinez, B. (2020) Personal communication.

Marx, J. D. and Cummings, K. (2007) 'Normalized change', *American Journal of Physics*, vol. 75.

Mathews, J. (2006) 'Just whose idea was all this testing?', *The Washington Post*, 14th November 2006.

Mazur, E. (1997) *Peer Instruction: A Users Manual*, Prentice-Hall, New Jersey.

McBride, W. (2009) *Teaching to Gender Differences: Boys will be Boys and Girls will be Girls*, World Books, Chicago, Illinois.

McCloskey, M. (1983) 'Intuitive Physics', *Scientific American*, vol. 248, no. 4, pp. 122-130.

Miesekes, M., Pado, U. (2019) 'Summarization Evaluation meets Short-Answer Grading', *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*, Turku, Finland, pp. 79-85.

Millar, J. (2005) *Engaging students with assessment feedback: What works? An FDTL5 Project literature review*, Oxford Brookes University, Oxford, United Kingdom.

Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2002) 'Towards robust computerised marking of free-text responses', *6th International CAA Conference*, Loughborough, United Kingdom.

Muggleton, S. and de Raedt, L. (1994) 'Inductive logic programming: Theory and methods', *Journal of Logic Programming*, vol. 19, pp. 629-679.

Muller, T. and Frauendiener, J. (2011) 'Studying null and time-like geodesics in the classroom', *European Journal of Physics*, vol. 32, pp. 747-759.

Nardi, A. and Ranieri, M. (2019) 'Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students performance, self-efficacy and satisfaction', *British Journal of Educational Technology*, vol. 50, no. 3.

Nickerson, J., Corter, J., Esche, S. and Chassapis, C. (2007) 'A model for evaluating the effectiveness of remote engineering laboratories and simulations in education', *Computers and Education*, vol. 49, no. 3, pp. 708-725.

Nicol, D. (2007) 'E-assessment by design: using multiple-choice tests to good effect', Journal of Further and Higher Education, vol. 31, no. 1, 53-64.

Noorbehbahani, F. and Karden, A. A. (2011) 'The automatic assessment of free text answers using a modified BLEU algorithm', *Computers and Education*, vol. 56, no. 2, pp. 337-345.

Norton, A. J. (2017). Personal communication.

Norton, A. J. (2018). Personal communication.

Nyquist, J. B. (2003) 'The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis', *PhD dissertation*, Vanderbilt University, United States.

Papineni, K., Roukas, S., Ward, T. and Zhu, W. (2001) 'BLEU: a method for automatic evaluation of machine translation', *Technical Report RC22176 (WO109-022)*, IBM Research Division.

Parker, M. A. (2019). Personal communication.

Pereira, F. C. N. and Grosz, B. J (1994) *Natural language processing*, MIT Press, Massachusetts.

Perez, D., Alfonseca, E. and Rodriguez, P. (2004a) 'Application of the BLEU method for evaluating free-text answers in an e-learning environment', *Language Resources and Evaluation Conference (LREC-2004)*, Lisbon, Portugal.

Perez, D., Alfonseca, E. and Rodriguez, P. (2004b) 'Upper bounds and extension of the BLEU algorithm applied to assessing student essays', *International Association for Educational Assessment Conference (IAEA-2004)*, Philadelphia, United States.

Perez, D., Gliozzo, A. Strapparava, C., Alfonseca, E., Rodriguez, P. and Magnini, B. (2005) 'Automatic assessment of students' free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis', *Proceedings of the 18th international Florida artificial intelligence research society conference (FLAIRS05)*, pp. 358-362.

Perez-Marin, D., Pascual-Nieto, I., Rodriguez, P. (2009) 'Computer-assisted assessment of free-text answers', *The knowledge engineering review*, vol. 24, no. 4, pp. 353-374.

Physport website (2018). Available at https://www..org/recommendations/Entry.cfm?ID=93334 (Accessed 13th April 2018).

Pineda, K. (2019). Personal communication.

Porter, L., Taylor, C. and Webb, K. (2014) 'Leveraging open source principles for flexible concept inventory development', *Proceedings of the 2014 conference on innovation technology in computer science education*, pp. 243-248.

Poutot, G. and Blandin, B. (2015) 'Exploration of Students' Misconceptions in Mechanics using the FCI', *American Journal of Educational Research*, vol. 3, no. 2, pp. 116-120.

Prather, J. P. (1985) 'Philosophical Examination of the Problem of Un-learning of Incorrect Science Concepts', *National Association for Research in Science Teaching*, ERIC document ED256570.

Pulman, S. and Sukkarieh, J. (2005) 'Automatic Short Answer Marking', *Proceedings of the 2nd Workshop on Building Educational Applications using NLP.*

Qin, Y., Wen, Q. and Wang, J. (2009) 'Automatic evaluation of translation quality using expanded N-gram co-occurrence', *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1-5.

Rebello, N. and Zollman, D. (2004) 'The effect of distractors on student performance on the force concept inventory', *American Journal of Physics*, vol. 72.

Reed-Rhoads, T. and Imbrie, P. K. (2008) 'Concept inventories in Engineering Education'.

Reeves, B. and Nass, C. (1996) *The media equation*, Stanford, California.

Richardson, C. T. and O'Shea, B. W. (2013) 'Assessing gender differences in response system questions for an introductory physics course', *American Journal of Physics*, vol. 81.

Robertson, A. E., Stanfield, A. C., Watt, J., Barry, F., Day, M., Cormack, M. and Melville, C. (2018) 'The experience and impact of anxiety in autistic adults: A thematic analysis', *Research in Autism Spectrum Disorders*, vol. 46, pp. 8-18.

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M. and Gould, R. R. (2009) 'The Astronomy and Space Science Concept Inventory: Development and Validation of Assessment Instruments Aligned with the K-12 National Science Standards', *Astronomy Education Review*, vol. 8, no. 1.

Sangwin, C. J. (2013) *Computer aided assessment of mathematics*, Oxford: Oxford University Press.

Sarrouti, M. and El Alaoui, S. O. (2020) 'SemBioNLQA: A scientific biomedical question answering system for retrieving exact and ideal answers to natural language questions', *Artificial Intelligence in Medicine*, vol. 102.

Scanlon, E., Colwell, C., Cooper, M. and Di Paolo, T. (2004) 'Remote experiments, reversioning and re-thinking science learning', *Computers and Education*, vol. 43, pp. 153-163.

Schutter, A. and Shamir, L. (2015) 'Galaxy morphology An unsupervised machine learning approach', *Astronomy and Computing*, vol. 12, pp. 60-66.

Scott, S. V. (2014) 'Practising what we preach: Towards a student-centred definition of feedback', *Teaching in Higher Education*, vol. 19, no. 1, pp. 49-57.

Scott, T. F., Schumayer, D. and Gray, A. R. (2012) 'Exploratory Factor Analysis of a Force Concept Inventory Data Set', *Physical Review Special Topics Physics Education Research*, vol. 8, no. 2.

Scott, T. F. and Schumayer, D. (2017) 'Conceptual coherence of non-Newtonian worldviews in Force Concept Inventory data', *Physical Review Physics Education Research*, vol. 13.

Scott, T. F. and Schumayer, D. (2018) 'Central distractors in the Force Concept Inventory', *Physical Review Physics Education Research*, vol. 14.

Scott, W. A. (1955) 'Reliability of content analysis: The case of nominal scale coding', *Public Opinion Quarterly*, vol. 19, pp. 321-325.

Sedrakyan, G., Malmberg, J., Verbert, K., Jarvela, S. and Kirschner, P. A. (2018) 'Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation', *Computers in human behavior*, in press.

Semon, M. D., Malin, S. and Wortel, S. (2009) 'Exploring the transition from special to general relativity', *American Journal of Physics*, vol. 77, pp. 434-438.

Serjeant, S. (2018). Personal communication.

Shuhidan, S., Hamilton, M. and DSouza, D. (2010) 'Instructors perspectives of multiple-choice questions in summative assessment for novice programmers', *Computer Science Education*, vol. 20, no. 3, pp. 229-259.

Shute, V. J. (2008) 'Focus on formative feedback', *Review of Educational Research*, vol. 78, no. 1, pp. 153-189.

Simon and Snowdon, S. (2014) 'Multiple-choice vs free-text code-explaining examination questions', *Proceedings of the 14th Koli Calling International Conference on computing education research*, pp. 91-97.

Smedley, R. M. and Coulson, N. S. (2017) 'A themaic analysis of messages posted by moderators within health-related asynchronous online support forums', *Patient Education and Counseling*, vol. 100, no. 9, pp. 1688-1693.

Smith, J. I. and Tanner, K. (2010) 'The problem of revealing how students think: Concept inventories and beyond', *CBE life sciences education*, vol. 9, no. 1, pp. 1-5.

Smith, T. I., Louis, K. J., Ricci, B. J. and Bendjilali, N. (2020) 'Quantitatively ranking incorrect responses to multiple-choice questions using item response theory', *Physical Review Physics Education Research*, vol. 16.

Srinivasan, A. (2004) *The Aleph Manual*. Available at https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html (Accessed 3rd July, 2019)

Stannard, W., Kersting, M., Kraus, U. and Moschilla, J. (2017) 'Research into the teaching and learning of Einsteinian physics in international contexts', *GIREP-ICPE-EPEC*, Dublin, 7th July 2017.

Steif, P. and Dantzler, J. (2005) 'A statics concept inventory: Development and psychometric analysis', *Journal of Engineering Education*, vol. 94, pp. 363-371.

Sterling, L. and Shapiro, E. (1994) *The Art of PROLOG: Advanced Programming Techniques*, MIT Press.

Stocklmayer, S., Rayner, J. P. and Gore, M. M. (2012) 'Changing the Order of Newton's Laws Why How the Third Law should be First', *The Physics Teacher*, vol. 50.

Stoen, S. M., McDaniel, M. A., Frey, R. F., Hynes, K. M. and Cahill, M. J. (2020) 'Force Concept Inventory: More than just conceptual understanding', *Physical Review Physics Education Research*, vol. 16.

Sukkarieh, J., Pulman, S. and Raikes, N. (2003) 'Auto-marking: using computational linguistics to score short, free-text responses', *29th International Association for Educational Assessment (IAEA) Annual Conference*, Manchester, United Kingdom.

Sychev, O., Anikin, A. and Prokudin, A. (2020) 'Automatic grading and hinting in open-ended text questions', *Cognitive Systems Research*, vol. 59, pp. 264-272.

The Open Science Laboratory website (2020). Available at https://learn5.open.ac.uk/course/view.php?id=2 (Accessed 27th July 2020).

Thornton, R. and Sokoloff, D. (1998) 'Assessing student learning of Newtons laws: The force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula', *American Journal of Physics*, vol. 66.

Traxler, A., Henderson, R., Stewart, J., Stewart, G., Papak, A. and Lindell, R. (2018) 'Gender fairness within the Force Concept Inventory', *Physical Review Physics Education Research*, vol. 14.

University of Auckland website (2017). Available at https://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis/about-thematic-analysis.html (Accessed 24th July 2017).

Varagona, L. M. and Hold, J. L. (2019) 'Nursing students perceptions of faculty trustworthiness: Thematic analysis of a longitudinal study', *Nurse Education Today*, vol. 72, pp. 27-31.

Walker, D. J., Topping, K. and Rodrigues, S. (2008) 'Student reflections on formative e-assessment: Expectations and perceptions', *Learning, Media Technology*, vol. 33, no. 3, pp. 221-234.

Wallace, C. S. and Bailey, J. M. (2010) 'Do Concept Inventories Actually Measure Anything?', *Astronomy Education Review*, vol. 9, no. 1.

Weaver, M. R. (2006) 'Do students value feedback? Student perceptions of tutors' written responses', *Assessment and Evaluation in Higher Education*, vol. 31, no. 3, pp. 379-394.

Wells, J., Henderson, R., Stewart, J., Stewart, G., Yang, J. and Traxler, A. (2019) 'Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis', *Physical Review Physics Education Research*, vol. 15.

White, B. Y. (1983) 'Sources of Difficulty in Understanding Newtonian Dynamics', *Cognitive Science*, vol. 7, pp. 41-65.

Willis, A. (2010) *Inductive Logic Programming to Support Automatic Marking of Student Answers in Free Text, Final Report*, The Open University, Milton Keynes, United Kingdom.

Woodford, K. and Bancroft, P. (2005) 'Multiple choice questions not considered harmful', *Seventh Australasian Computing Education Conference (ACE 2005)*, pp. 109-115.

Yasuda, J., Mae, N., Hull, M. M. and Taniguchi, M. (2018) 'Analyzing false positives of four questions in the Force Concept Inventory', *Physical Review Physics Education Research*, vol. 14.

Yeo, S. and Zadnik, M. (2000) 'Newton, we have a problem...', *Australian Science Teacher*, vol. 46.

Zahn, C. and Kraus, U. (2014) 'A toolkit for teaching general relativity: I. Curved spaces and spacetimes', *European Journal of Physics*, vol. 35.

Zehner, F., Salzer, C. and Goldhammer, F. (2016) 'Automatic Coding of Short Text Responses via clustering in Educational Assessment', *Educational and Psychological Measurement*, vol. 76, no. 2, pp. 280-303.

Zhu, M., Liu, O. L. and Lee, H. S. (2020) 'The effect of automated feedback on revision behaviour and learning gains in formative assessment of scientific argument writing', *Computers and Education*, vol. 143.

Zilvinskis, J., Willis, J. and Bordon, V. M. H. (2017) 'An Overview of Learning Analytics', *New Directions for Higher Education*, vol. 2017, pp. 9-17.

Zwick, R. (1988) 'Another Look at Interrater Agreement', *Psychological Bulletin*, vol. 103, no. 3, pp. 374-378.

12 Appendix A: AMS Questions and marking rules

This appendix gives tables which show how the questions on the different versions of the AMS map to one another, and to the questions on the original version of the FCI. In addition, it also gives the questions and marking rules from the final version of the AMS.

The non-trivial table headings and abbreviations used in Tables A.1 and A.2 are explained as follows. Note that here, **emboldening** does not follow the convention of the rest of the thesis; it is used for emphasis, rather than referring to *Glossary* terms found in **Appendix I**.

• **Standardized AMS question** refers to the question numbering from Version 1 of the AMS, and this numbering is taken as the standardized question numbering throughout the different versions; this numbering is employed to avoid confusion when discussing different versions of the same question from other versions of the AMS.

• **Question type (VX)** refers to the question type in a particular version of the AMS. For example, *Question type (V1)* would indicate the question type in Version 1 of the AMS.

• **FRQ** is the abbreviation for *free-response question*; **MCQ** is the abbreviation for *multiple-choice question*; and **MRQ** is the abbreviation for *multiple-response question.*

• **FRQ(L)** is the abbreviation for *free-response question (letter).* This is a free-response question that requires the entry of a single letter corresponding to a multiple-choice option. These types of question are often based on trajectories.

• **A blank entry -** indicates that a question was withdrawn in a particular version because of previous problems. Note that the only two questions removed from the final version were extensions to the original FCI.

Standardized AMS question	FCI question	Question type (V1)	Question type (V2)	Question type (V3)	Question type (V4)	Situation	Concept Tested
Q1	Q1	FRQ	FRQ	FRQ	FRQ	Balls on table	Newton's second law
Q2	Q2	FRQ	FRQ	FRQ	FRQ	Balls on table	Kinematics
Q3	Q3	FRQ	FRQ	FRQ	FRQ	Ball drop	Kinematics (Projectile motion)
Q4	Q3	FRQ	FRQ	FRQ	-	Ball drop	Kinematics
Q5	Q4	FRQ	FRQ	FRQ	FRQ	Truck and car	Newton's Third Law
Q6	Q5	MRQ	-	-	MRQ	Marble in track	Types of Forces
Q7	Q5	MRQ	FRQ	FRQ	FRQ	Marble in track	Types of Forces
Q8	Q6	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Marble in track	Newton's First Law (Circular motion)
Q9	Q7	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Hammer throw	Newton's First Law (Circular motion)
Q10	Q8	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Hockey puck	Newton's second law
Q11	Q9	FRQ	FRQ	FRQ	FRQ	Hockey puck	Kinematics (Collisions)
Q12	Q10	MCQ	FRQ	FRQ	FRQ	Hockey puck	Newton's First Law
Q13	Q11	FRQ	FRQ	FRQ	FRQ	Hockey puck	Types of Forces
Q14	Q12	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Cannon	Kinematics (Projectile motion)
Q15	Q13	MCQ	FRQ	FRQ	FRQ	Ball toss	Kinematics (Projectile motion)
Q16	Q14	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Bowling ball	Kinematics (Projectile motion)
Q17	Q15	FRQ	FRQ	FRQ	FRQ	Truck and car	Newton's Third Law
Q18	Q16	MCQ	FRQ	FRQ	FRQ	Truck and car	Newton's Third Law

Table A.1: Table showing how the questions on different versions of the AMS map to one another, and to the original FCI questions.

Standardized AMS question	FCI question	Question type (V1)	Question type (V2)	Question type (V3)	Question type (V4)	Situation	Concept Tested
Q19	Q17	FRQ	FRQ	FRQ	-	Elevator	Types of Forces
Q20	Q17	FRQ	FRQ	FRQ	FRQ	Elevator	Newton's First Law
Q21	Q18	MRQ	FRQ	FRQ	FRQ	Boy on swing	Types of Forces
Q22	Q19	FRQ	FRQ	FRQ	FRQ	Moving blocks	Kinematics (Velocity)
Q23	Q20	FRQ	FRQ	FRQ	FRQ	Moving blocks	Kinematics (Acceleration)
Q24	Q21	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Rocket	Newton's second law
Q25	Q22	FRQ	FRQ	FRQ	FRQ	Rocket	Newton's First Law
Q26	Q23	MCQ	FRQ(L)	FRQ(L)	FRQ(L)	Rocket	Kinematics
Q27	Q24	FRQ	FRQ	FRQ	FRQ	Rocket	Newton's First Law
Q28	Q25	MCQ	FRQ	FRQ	FRQ	Woman pushing box	Newton's First Law
Q29	Q26	FRQ	FRQ	FRQ	FRQ	Woman pushing box	Newton's second law
Q30	Q27	FRQ	FRQ	FRQ	FRQ(L)	Woman pushing box	Kinematics
Q31	Q28	FRQ	FRQ	FRQ	FRQ	Students in chairs	Newton's Third Law
Q32	Q29	FRQ	FRQ	FRQ	FRQ	Chair on floor	Newton's Third Law
Q33	Q30	MRQ	FRQ	FRQ	MRQ	Tennis ball	Types of Forces

Table A.2: Table showing how the questions on different versions of the AMS map to one another, and to the original FCI questions.

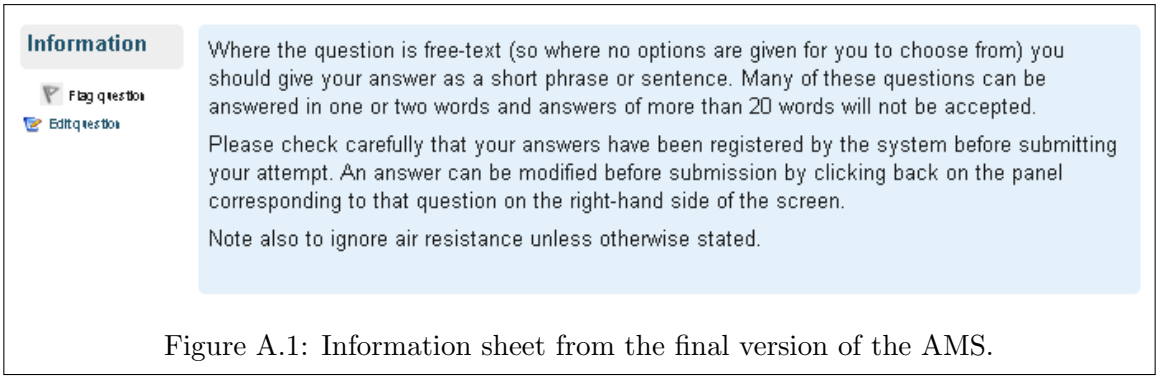


Figure A.1: Information sheet from the final version of the AMS.

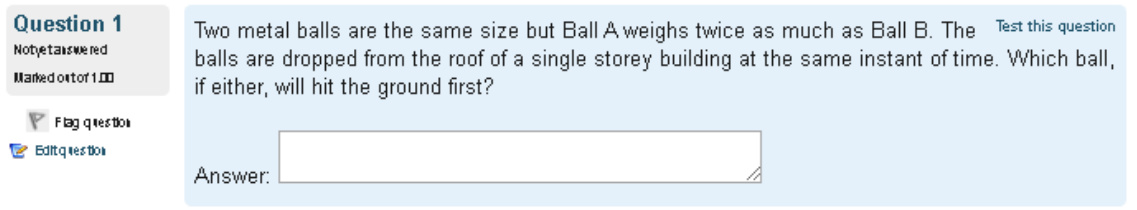


Figure A.2: Question 1 from the final version of the AMS.

```
Q1

match_any (
    match_mw (dont know)
    match_w (either)
    match_w (one)
) #Negative

match_any (
    match_mw (same)
    match_mw (simultaneous)
    match_mw (simultaneously)
    match_mw (neither)
    match_mw (none)
    match_mw (both)
    match_mw (identical)
    match_mw (together)
) #Positive
```

Figure A.3: Marking rules for Question 1 of the final version of the AMS.

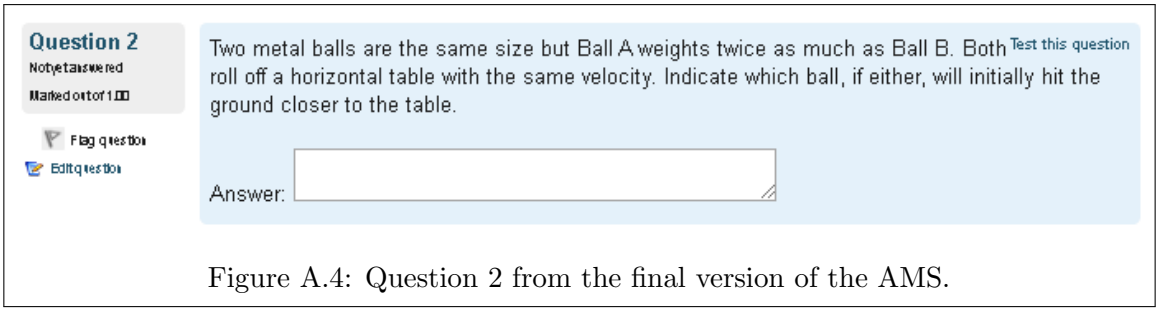


Figure A.4: Question 2 from the final version of the AMS.

```
Q2

match_mw (dont know) #Negative

match_any (
    match_mw (neither closer)
    match_mw (equal* close*)
) #Positive

match_any (
    match_mw (further)
    match_mw (nearer)
    match_mw (closer)
    match_mw (twice)
    match_mw (half)
    match_w (one)
    match_w (either)
) #Negative

match_any (
    match_mw (same)
    match_mw (neither)
    match_mw (equal)
    match_mw (equally)
    match_mw (both)
    match_mw (none)
) #Positive
```

Figure A.5: Marking rules for Question 2 of the final version of the AMS.

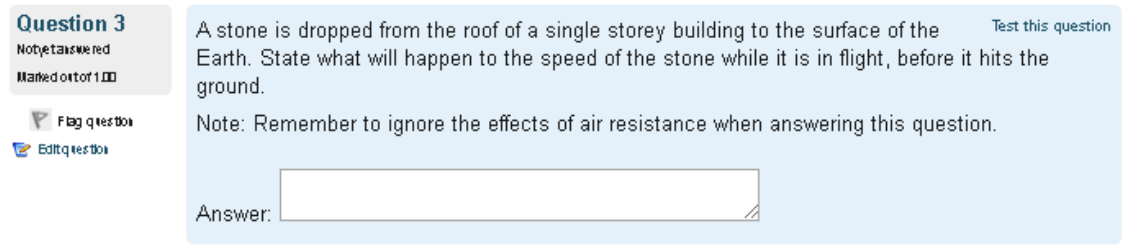


Figure A.6: Question 3 from the final version of the AMS.

```
Q3

match_any (
    match_mw (dont know)
    match_mw (terminal)
    match_mw (max*)
) #Negative

match_any (
    match_mw (up)
    match_mw (increase*)
    match_mw (accelerat*)
    match_mw (faster)
    match_mw (great*)
    match_mw (rise)
) #Positive
```

Figure A.7: Marking rules for Question 3 of the final version of the AMS.

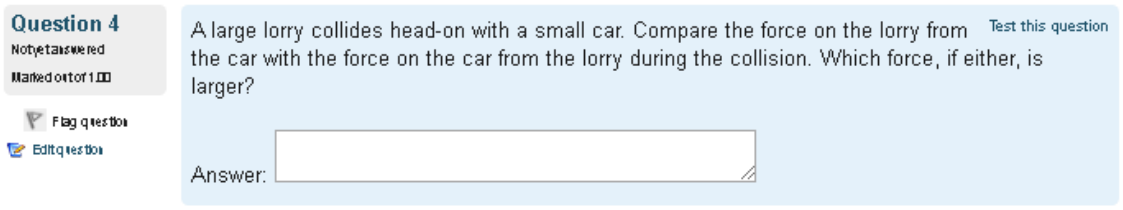


Figure A.8: Question 4 from the final version of the AMS.

```
Q4

match_any (
    match_mw (dont know)
    match_mw (not equal)
) #Negative

match_mw (neither greater*|larg*|higher*|bigger) #Positive

match_any (
    match_w (either)
    match_mw (greater*|larg*|higher*|bigger)
) #Negative

match_any (
    match_mw (neither)
    match_mw (same)
    match_mw (equal)
    match_mw (equivalent)
) #Positive
```

Figure A.9: Marking rules for Question 4 of the final version of the AMS.

Question 5

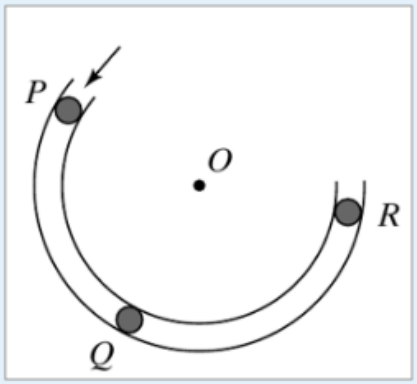
Not yet answered

Marked out of 1.00

Flag question

Edit question

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high speed into the channel at *P* and exits at *R*.



Which of the following forces are acting on the ball when it is in the frictionless channel at point *Q*?

Select one or more:

- A downward force of gravity.
- A force pointing from Q to O.
- A force in the direction of motion.
- A force pointing from O to Q.
- An upward force from the table.

Figure A.10: Question 5 from the final version of the AMS.

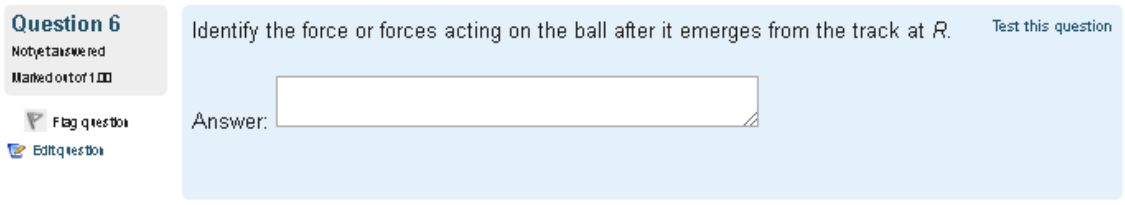


Figure A.11: Question 6 from the final version of the AMS.

```
Q6

match_any (
    match_mw (dont know)
    match_mow (motion)
    match_mw (momentum)
    match_mw (centrifug*)
    match_mw (thrust)
) #Negative

match_any (
    match_mow (gravity*|weight*|gravitational*|gravitation reaction)
    match_mow (gravity*|weight*|gravitational*|gravitation normal)
    match_mow (gravity*|weight*|gravitational*|gravitation surface)
    match_mow (gravity*|weight*|gravitational*|gravitation table)
    match_mow (gravity*|weight*|gravitational*|gravitation support)
    match_mow (gravity*|weight*|gravitational*|gravitation contact)
    match_mow (gravity*|weight*|gravitational*|gravitation upwards)
) #Positive
```

Figure A.12: Marking rules for Question 6 of the final version of the AMS.

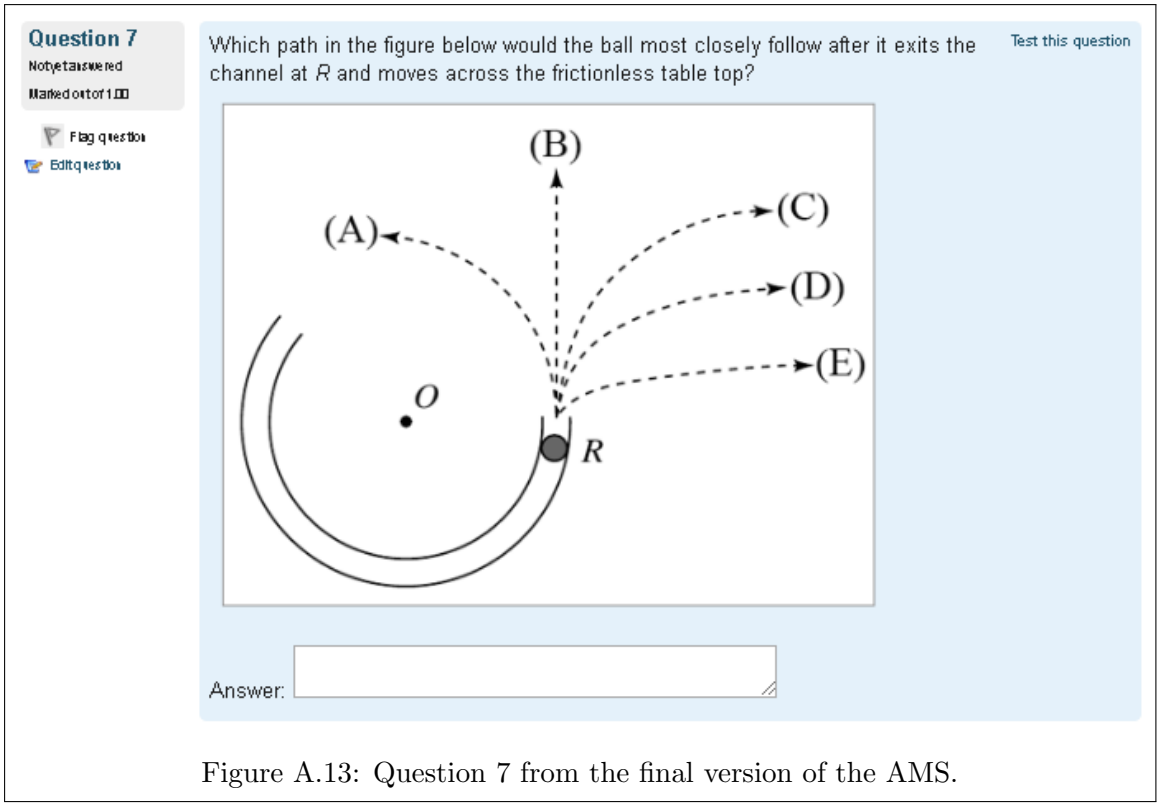


Figure A.13: Question 7 from the final version of the AMS.

```
Q7

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (B) #Positive
```

Figure A.14: Marking rules for Question 7 of the final version of the AMS.

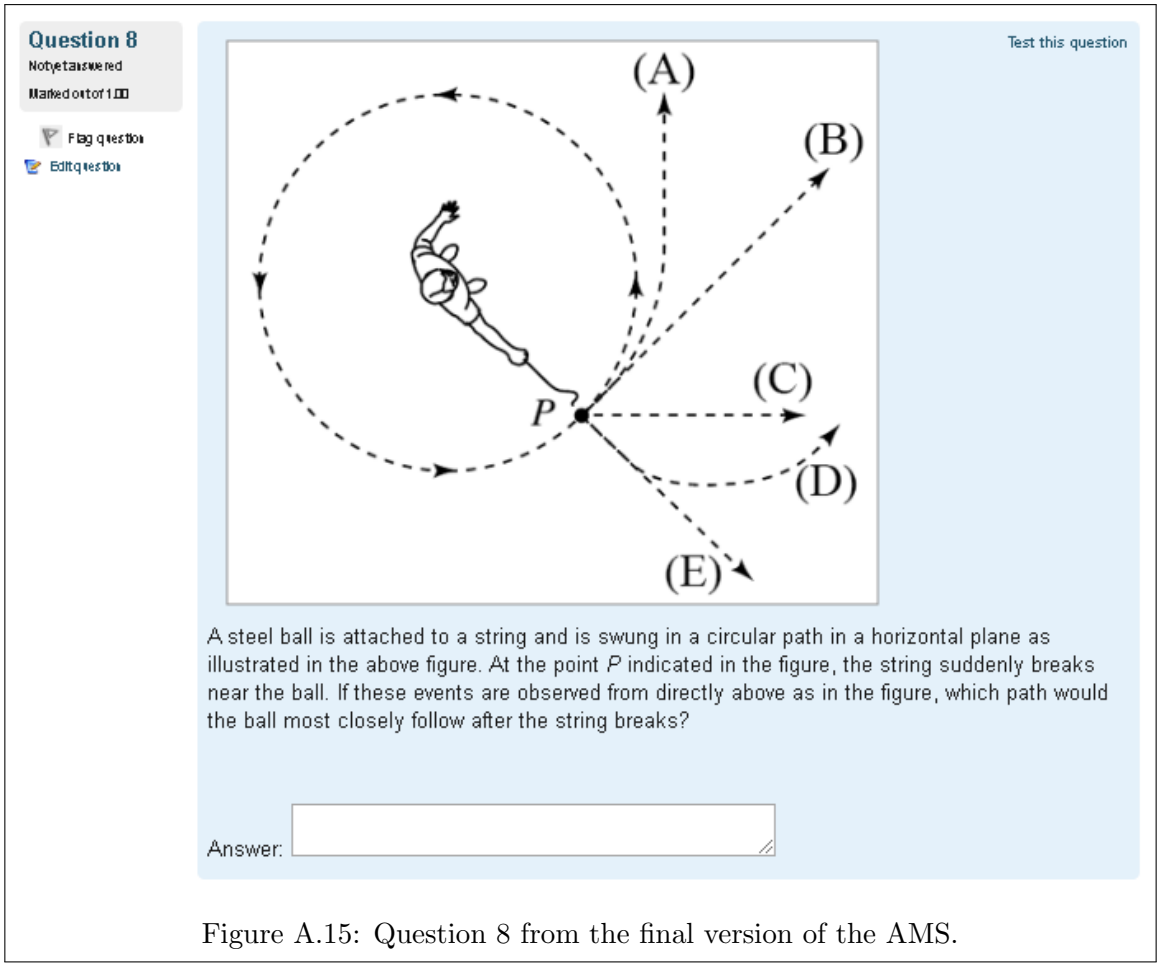


Figure A.15: Question 8 from the final version of the AMS.

```
Q8

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (B) #Positive
```

Figure A.16: Marking rules for Question 8 of the final version of the AMS.

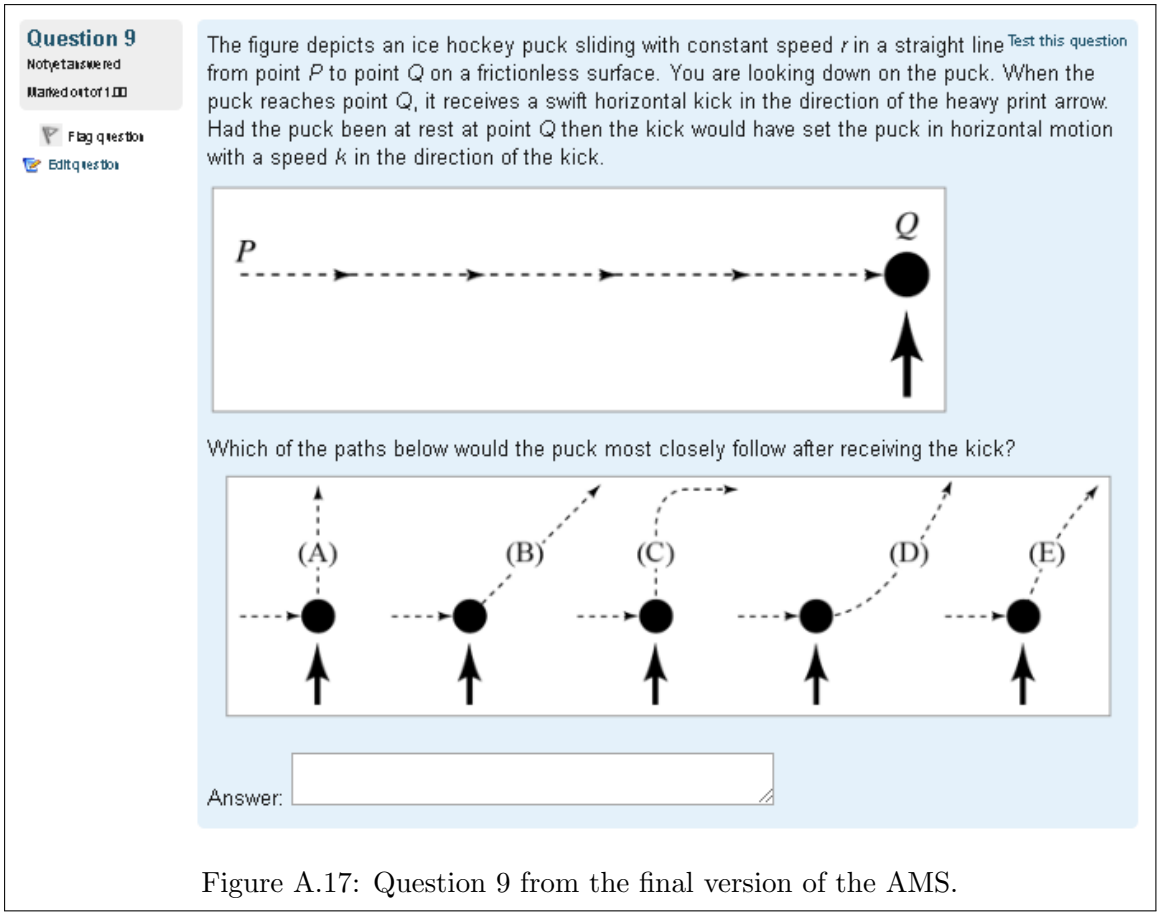


Figure A.17: Question 9 from the final version of the AMS.

```
Q9

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (B) #Positive
```

Figure A.18: Marking rules for Question 9 of the final version of the AMS.

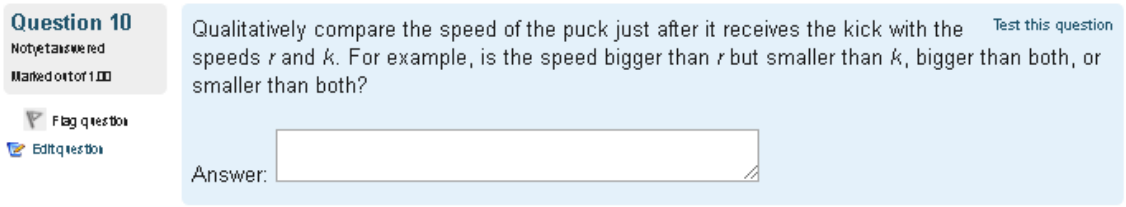


Figure A.19: Question 10 from the final version of the AMS.

```
Q10

match_mw (square root) #Positive

match_any (
    match_mw (dont know)
    match_mw (smaller)
    match_mw (same*|equal*|unchanged)
) #Negative

match_any (
    match_mw (bigger*|greater*|larger*|higher both)
    match_mw (bigger*|greater*|larger*|higher than r and k)
    match_mw (bigger*|greater*|larger*|higher than k and r)
) #Positive
```

Figure A.20: Marking rules for Question 10 of the final version of the AMS.

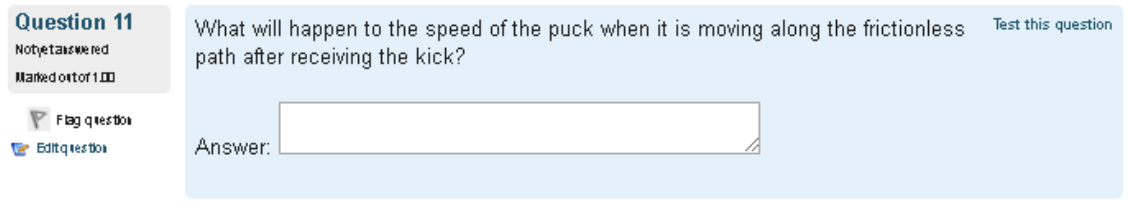


Figure A.21: Question 11 from the final version of the AMS.

```
Q11

match_any (
    match_mw (dont know)
    match_mw (increas*)
    match_mw (decreas*)
    match_mw (accelerat*)
    match_mw (either)
) #Negative

match_any (
    match_mw (constant*)
    match_mw (same*|remain*|maintain*)
    match_mw (no change)
    match_mw (not change)
    match_mw (unchanged)
    match_mw (nothing)
) #Positive
```

Figure A.22: Marking rules for Question 11 of the final version of the AMS.

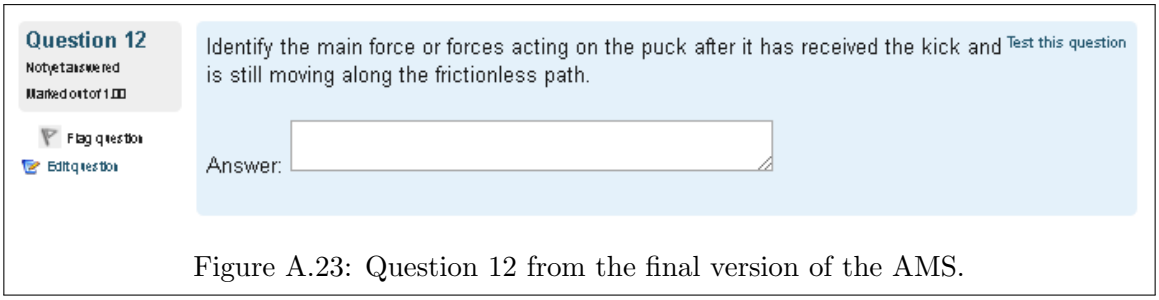


Figure A.23: Question 12 from the final version of the AMS.

```
Q12

match_any (
    match_mw (dont know)
    match_mow (kick*)
    match_mow (motion)
    match_mw (momentum)
    match_mw (driv*)
    match_mw (thrust)
) #Negative

match_any (
    match_mow (gravity*|weight*|gravitational*|gravitation reaction)
    match_mow (gravity*|weight*|gravitational*|gravitation normal)
    match_mow (gravity*|weight*|gravitational*|gravitation surface)
    match_mow (gravity*|weight*|gravitational*|gravitation ice)
    match_mow (gravity*|weight*|gravitational*|gravitation support)
    match_mow (gravity*|weight*|gravitational*|gravitation contact)
    match_mow (gravity*|weight*|gravitational*|gravitation upwards)
) #Positive
```

Figure A.24: Marking rules for Question 12 of the final version of the AMS.

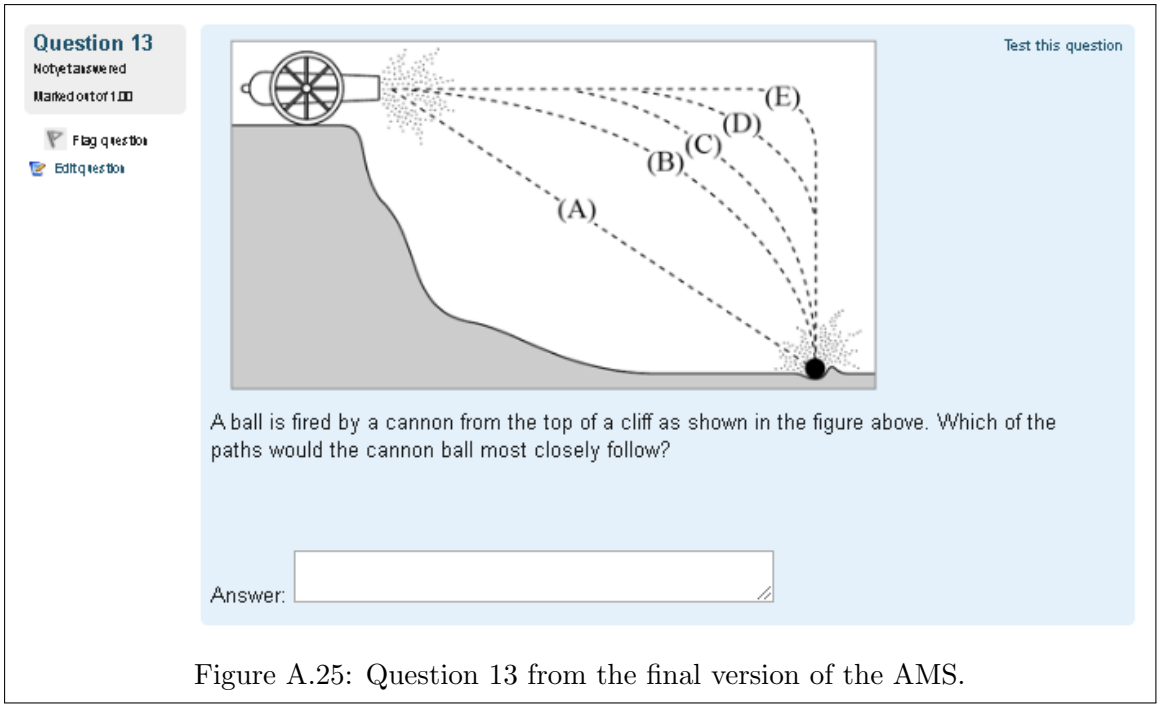


Figure A.25: Question 13 from the final version of the AMS.

```
Q13

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (B) #Positive
```

Figure A.26: Marking rules for Question 13 of the final version of the AMS.

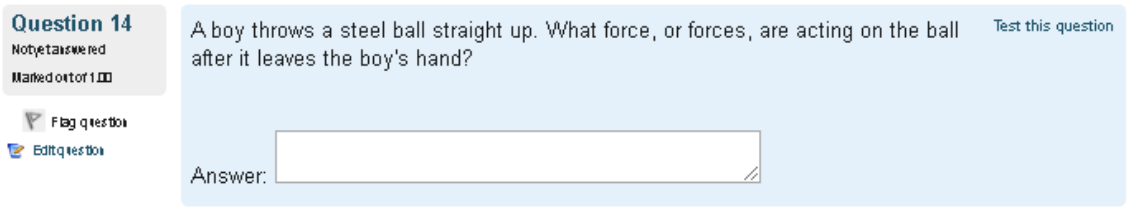


Figure A.27: Question 14 from the final version of the AMS.

```
Q14

match_any (
    match_mw (dont know)
    match_mw (electrostatic)
    match_mw (pressure)
    match_mw (internal)
    match_mw (kinetic)
    match_mw (energy)
    match_mw (push*)
    match_mw (mass)
    match_mw (thrust)
    match_mw (reaction)
    match_mw (initial force)
    match_mow (up* force)
    match_mw (lift)
    match_mw (throw*)
    match_mw (upthrust)
    match_mw (tension)
    match_mw (resultant force)
    match_mw (applied force)
    match_mowp0 (gravity and weight)
    match_mowp0 (weight and gravity)
    match_mw (drive)
    match_mw (potential)
    match_mw (friction)
) #Negative

match_any (
    match_mw (gravity)
    match_mw (gravitational)
    match_mw (weight)
) #Positive
```

Figure A.28: Marking rules for Question 14 of the final version of the AMS.

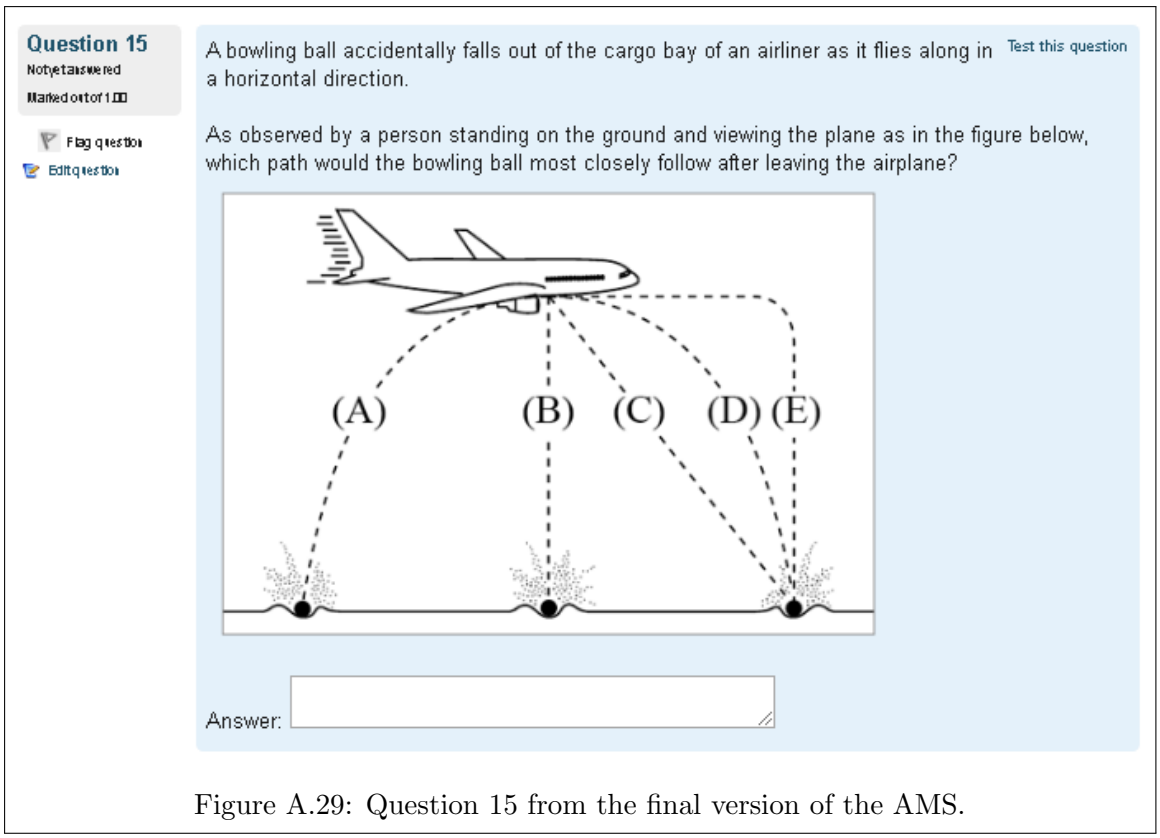


Figure A.29: Question 15 from the final version of the AMS.

```
Q15

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (D) #Positive
```

Figure A.30: Marking rules for Question 15 of the final version of the AMS.

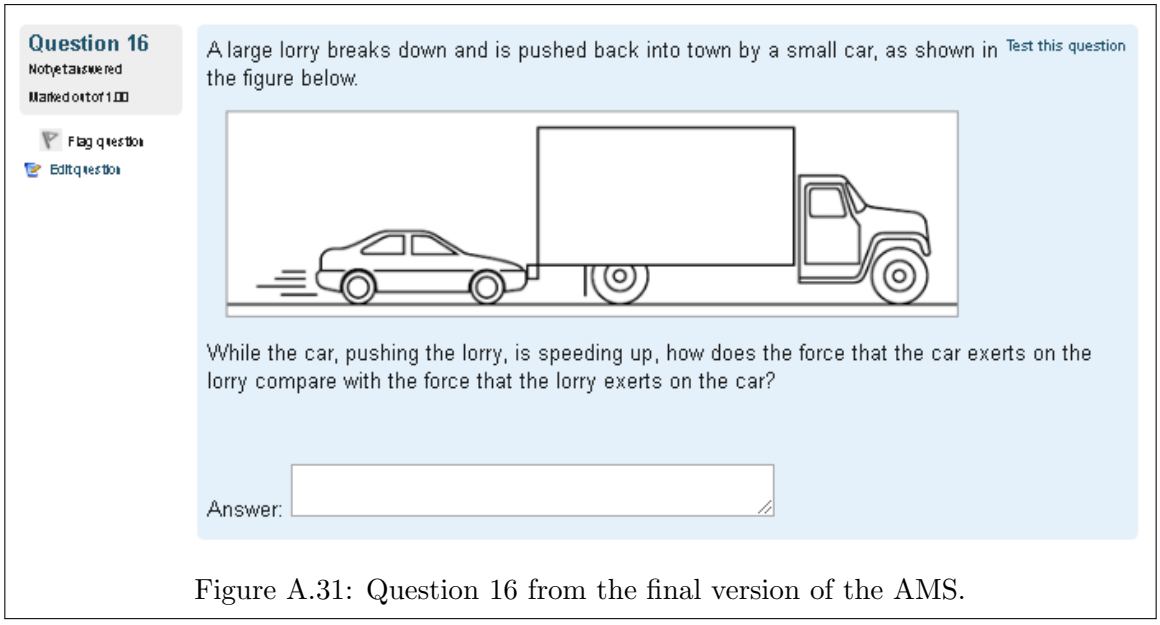


Figure A.31: Question 16 from the final version of the AMS.

```
Q16

match_any (
    match_mw (dont know)
    match_mw (greater*|larger*|increas*)
) #Negative

match_mw (same*|equal*|identical) #Positive
```

Figure A.32: Marking rules for Question 16 of the final version of the AMS.

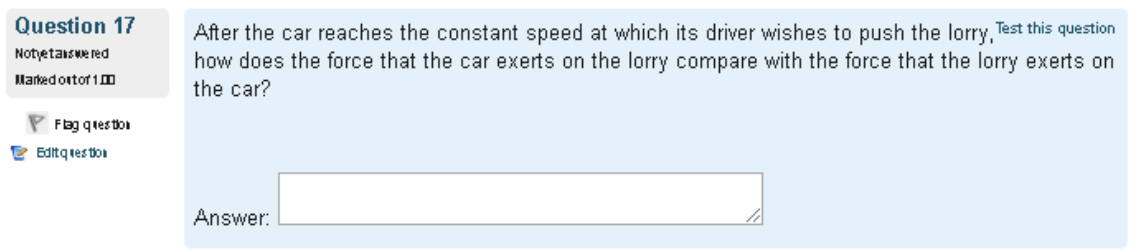


Figure A.33: Question 17 from the final version of the AMS.

```
Q17

match_any (
    match_mw (dont know)
    match_mw (greater*|larger*|increas*)
) #Negative

match_mw (same*|equal*|balanc*|identical) #Positive
```

Figure A.34: Marking rules for Question 17 of the final version of the AMS.

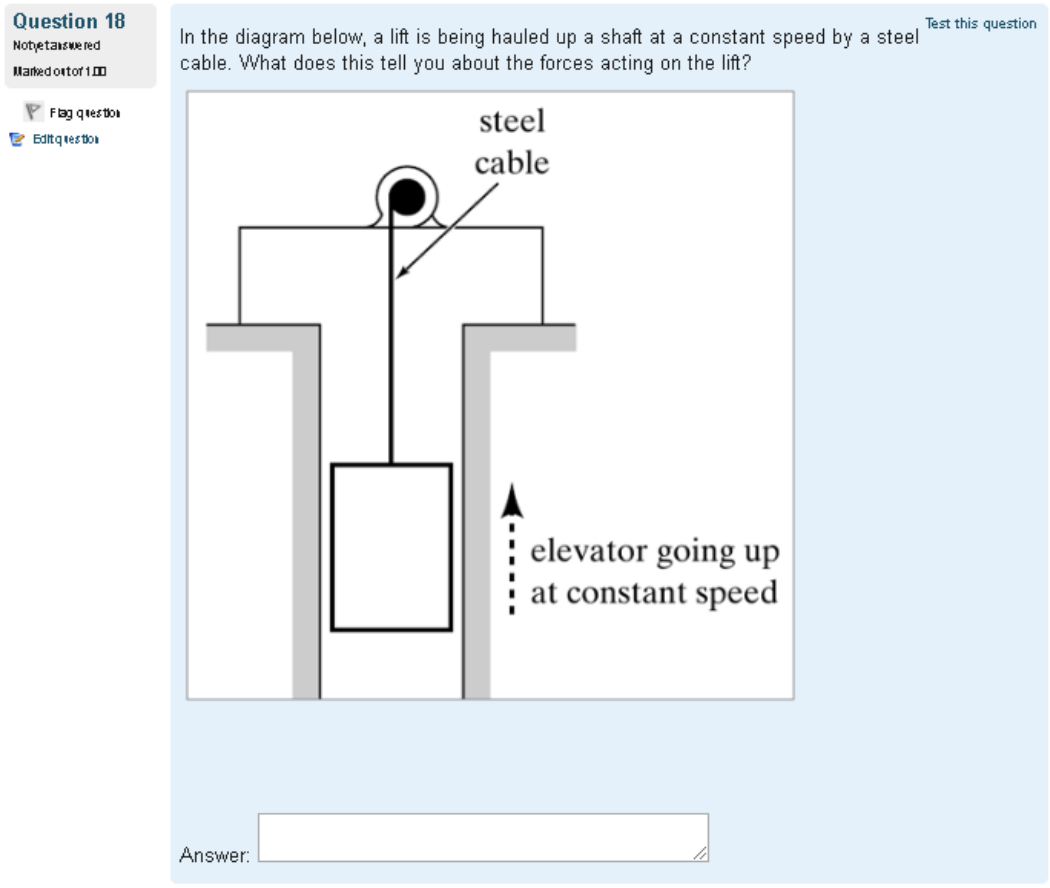


Figure A.35: Question 18 from the final version of the AMS.

```
Q18

match_any (
    match_mw (dont know)
    match_mw (greater)
    match_mw (smaller)
    match_mw (less)
    match_mw (unbalanced)
    match_mwp0 (not balanced)
) #Negative

match_any (
    match_mw (no net force)
    match_mw (no resultant force)
    match_mw (balance*|balanced)
    match_mw (equal*|equilibrium*|equivalent)
    match_mw (zero*|0*|none*|nothing*|null)
    match_mw (same*|identical)
    match_mw (cancel)
) #Positive
```

Figure A.36: Marking rules for Question 18 of the final version of the AMS.

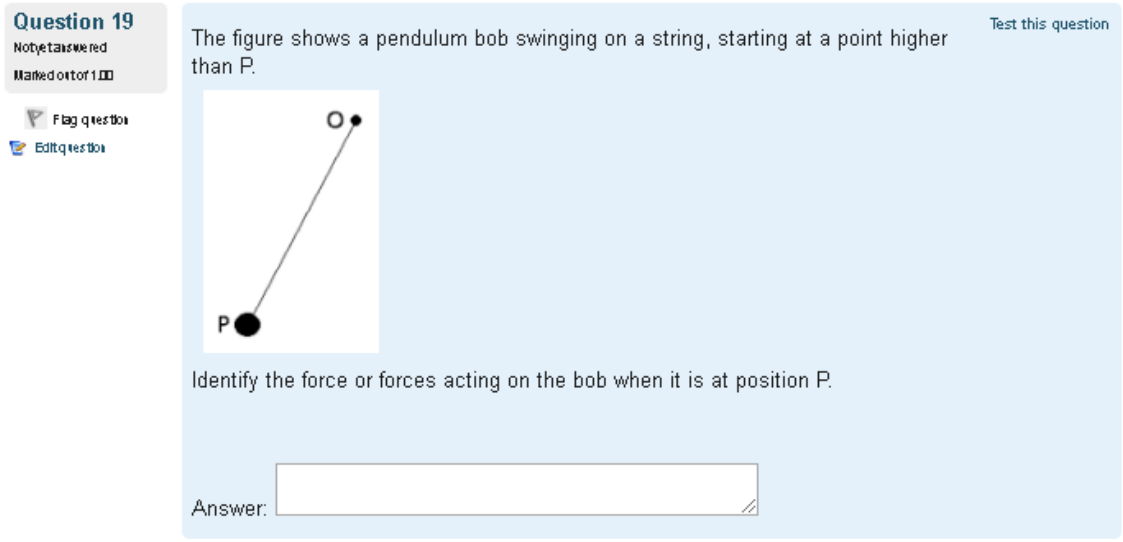


Figure A.37: Question 19 from the final version of the AMS.

```
Q19

match_any (
    match_mw (dont know)
    match_mw (thrust)
    match_mw (applied)
    match_mw (accel*)
    match_mw (centri*)
) #Negative

match_any (
    match_mow (gravit* tension)
    match_mow (weight tension)
    match_mow (gravit* force string)
    match_mow (weight force string)
    match_mow (gravit* pull* string)
    match_mow (weight pull* string)
) #Positive
```

Figure A.38: Marking rules for Question 19 of the final version of the AMS.

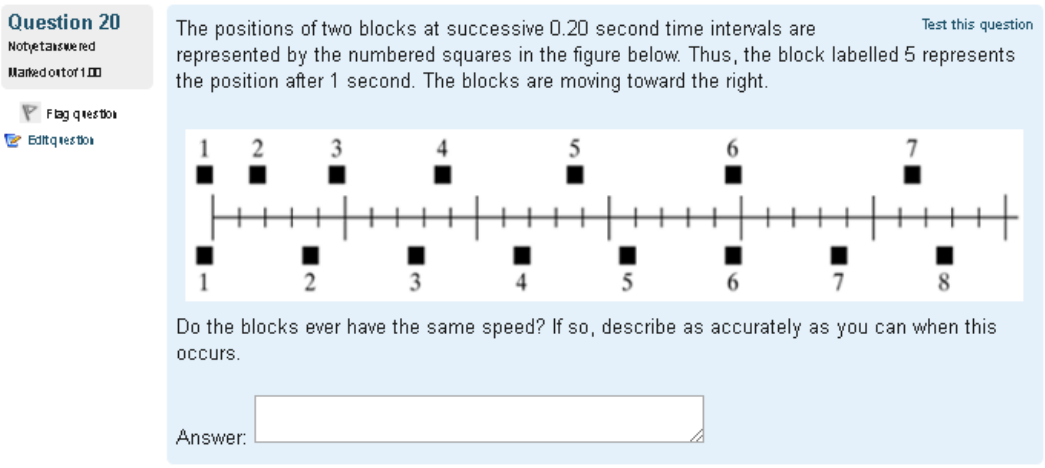


Figure A.39: Question 20 from the final version of the AMS.

```
Q20

match_any (
    match_mw (dont know)
    match_mw (0.4)
    match_mw (1.2)
    match_mw (1)
    match_mw (2)
    match_mw (5)
) #Negative

match_any (
    match_mw (3)
    match_mw (4)
    match_mw (0.6*)
    match_mw (0.7*)
    match_mw (0.8)
    match_mw (0.6s)
    match_mw (0.7s)
    match_mw (0.8s)
    match_mw (3*)
    match_mw (4th)
    match_mw (third)
    match_mw (fourth)
    match_w (three)
    match_mw (four)
) #Positive
```

Figure A.40: Marking rules for Question 20 of the final version of the AMS.

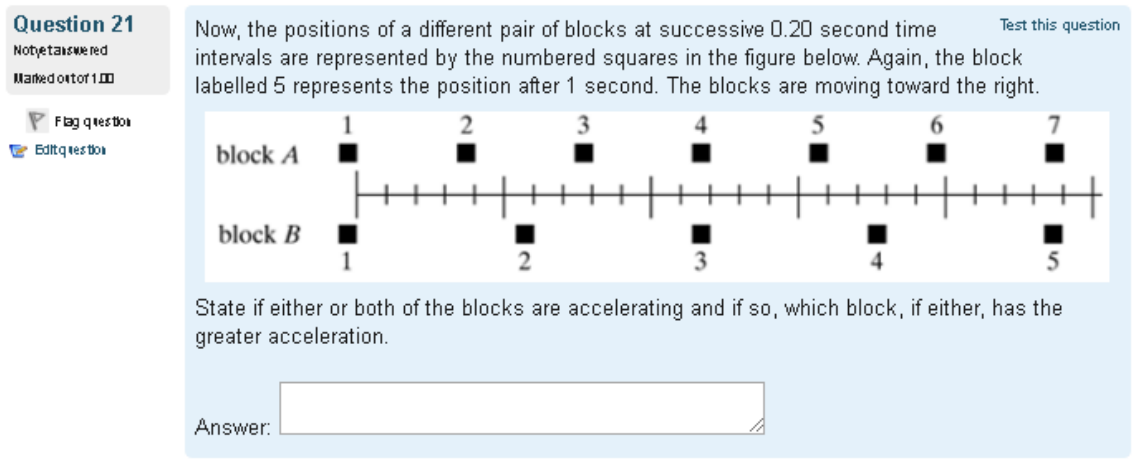


Figure A.41: Question 21 from the final version of the AMS.

```
Q21

match_mw (dont know) #Negative

match_any (
    match_mow (neither one)
    match_mw (neither greater)
) #Positive

match_any (
    match_w (one)
    match_w (greater)
    match_mw (both accelerat*)
    match_mw (deccerlat*)
    match_mw (decelerat*)
) #Negative

match_any (
    match_mw (neither*|both*|zero*|none)
    match_w (no)
    match_mw (not accelerat*)
    match_mwp0 (constant speed)
) #Positive
```

Figure A.42: Marking rules for Question 21 of the final version of the AMS.

Question 22

Not yet answered

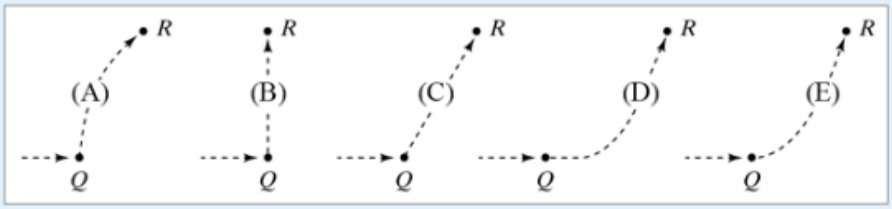
Marked out of 1.00

Flag question

Edit question

A rocket drifts sideways in outer space from point *P* to point *Q* as shown below. The rocket is subject to no outside forces. Starting at position Q, the rocket's engine is turned on and immediately produces a constant thrust (force on the rocket) at right angles to the line *PQ*. The constant thrust is maintained until the rocket reaches a point *R* in space (not shown). Test this question

Which of the paths below best represents the path of the rocket between points *Q* and *R*?



Answer:

Figure A.43: Question 22 from the final version of the AMS.

```
Q22

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (E) #Positive
```

Figure A.44: Marking rules for Question 22 of the final version of the AMS.

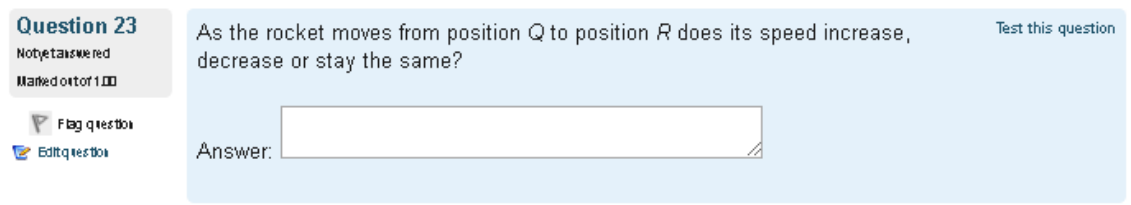


Figure A.45: Question 23 from the final version of the AMS.

```
Q23

match_any (
    match_mw (dont know)
    match_mw (same)
    match_mw (air)
    match_mw (decreas*)
    match_mw (constant)
) #Negative

match_any (
    match_mw (increas*)
    match_mw (accelerat*)
    match_mw (speed* up)
) #Positive
```

Figure A.46: Marking rules for Question 23 of the final version of the AMS.

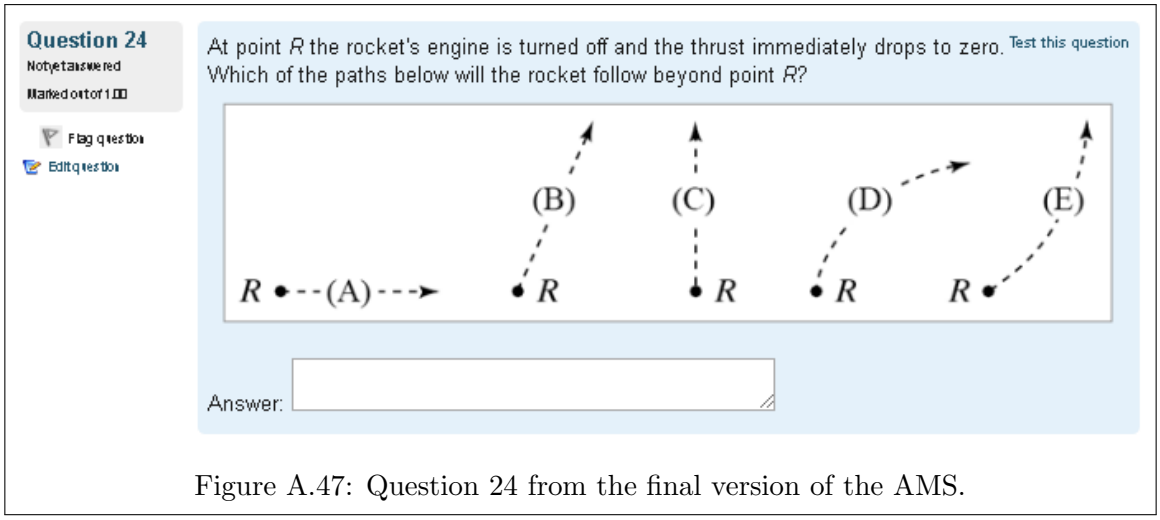


Figure A.47: Question 24 from the final version of the AMS.

```
Q24

match_any (
    match_mw (dont know)
    match_mw (or)
) #Negative

match_mw (B) #Positive
```

Figure A.48: Marking rules for Question 24 of the final version of the AMS.

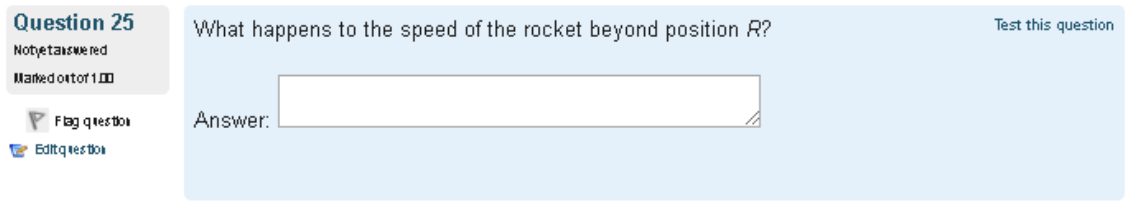


Figure A.49: Question 25 from the final version of the AMS.

```
Q25

match_any (
    match_mw (dont know)
    match_mw (increas*)
    match_mw (decreas*)
    match_mw (decelerat*)
) #Negative

match_any (
    match_mw (constant*)
    match_mw (same*|remain*|maintain*)
    match_mw (no change)
    match_mw (not change)
    match_mw (unchanged)
) #Positive
```

Figure A.50: Marking rules for Question 25 of the final version of the AMS.

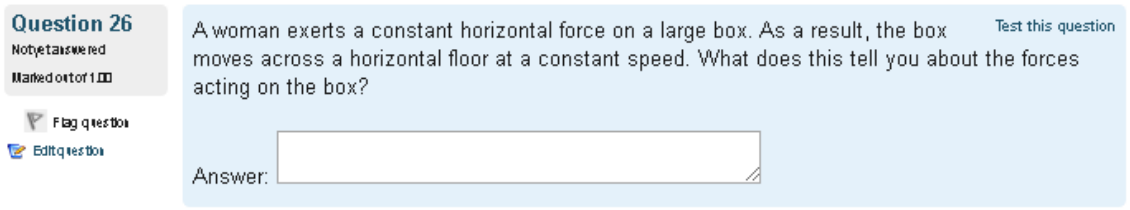


Figure A.51: Question 26 from the final version of the AMS.

```
Q26

match_any (
    match_mw (dont know)
    match_mw (greater)
    match_mw (smaller)
    match_mw (less)
    match_mwp0 (not balanced)
    match_mwp0 (not equal)
    match_mw (unbalanced)
    match_mw (gravity)
    match_mw (reaction)
) #Negative

match_any (
    match_mw (no net force)
    match_mw (no resultant force)
    match_mw (balance*|balanced)
    match_mw (equal*|equilibrium*|equivalent)
    match_mw (zero*|0*|none*|nothing*|null)
    match_mw (same*|identical)
    match_mw (cancel)
) #Positive
```

Figure A.52: Marking rules for Question 26 of the final version of the AMS.

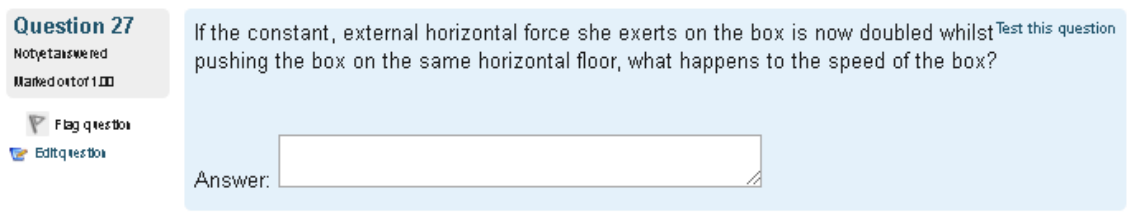


Figure A.53: Question 27 from the final version of the AMS.

```
Q27

match_mw (dont know) #Negative

match_mw (not doubl*) #Positive

match_any (
    match_mw (doubl*|two*|twice*|2)
    match_mw (root)
) #Negative

match_any (
    match_mw (increas*)
    match_mw (accelerat*)
    match_mw (up)
    match_mw (rise|rises)
    match_mw (faster)
    match_mw (higher)
    match_mw (bigger)
    match_mw (greater)
) #Positive
```

Figure A.54: Marking rules for Question 27 of the final version of the AMS.

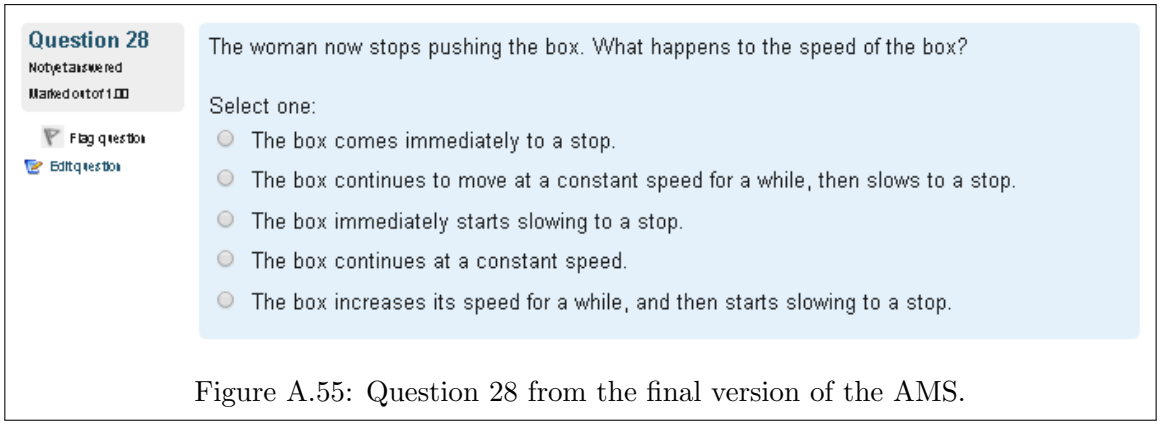


Figure A.55: Question 28 from the final version of the AMS.

Question 29

Not yet answered

Marked out of 1.00

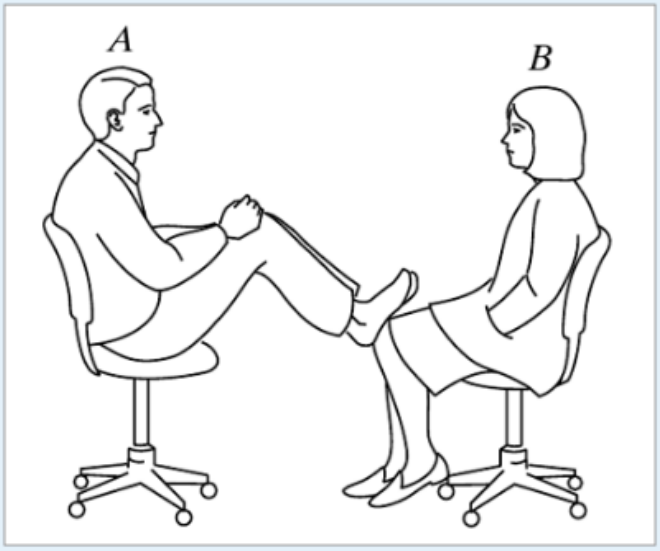
Flag question

Edit question

In the figure below, student *A* has a mass of 95 kg and student *B* has a mass of 77 kg. They sit in identical office chairs facing each other. Student *A* places his feet on the knees of student *B*, as shown. Student *A* then suddenly pushes outward with his feet, causing both chairs to move.

Test this question

What can you say about the amount of force each student exerts on the other during the push?



Answer:

Figure A.56: Question 29 from the final version of the AMS.

```
Q29

match_any (
    match_mw (dont know)
    match_mw (greater*|larger)
    match_mw (smaller|less)
) #Negative

match_mw (same*|equal*|identical*|balanc*|matched) #Positive
```

Figure A.57: Marking rules for Question 29 of the final version of the AMS.

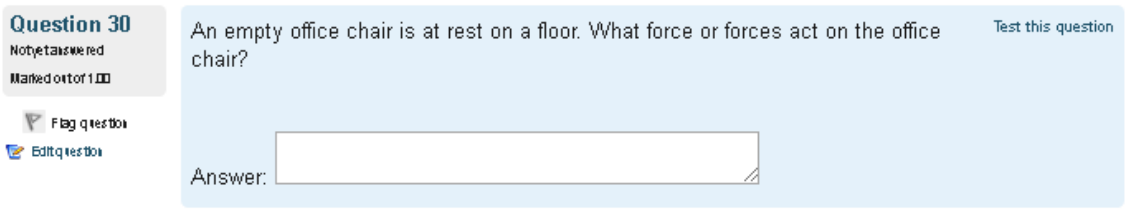


Figure A.58: Question 30 from the final version of the AMS.

```
Q30

match_mw (dont know) #Negative

match_any (
    match_mow (gravit* react*)
    match_mow (weight react*)
    match_mow (gravit* normal)
    match_mow (weight normal)
    match_mow (gravit* contact)
    match_mow (weight contact)
    match_mow (gravit* push* floor*|ground*|surface)
    match_mow (weight push* floor*|ground*|surface)
    match_mw (gravit* force floor*|ground*|surface)
    match_mw (force floor*|ground*|surface gravit*)
    match_mow (weight force floor*|ground*|surface)
    match_mow (weight resistance floor*|ground*|surface)
    match_mow (gravit* resistance floor*|ground*|surface)
) #Positive
```

Figure A.59: Marking rules for Question 30 of the final version of the AMS.

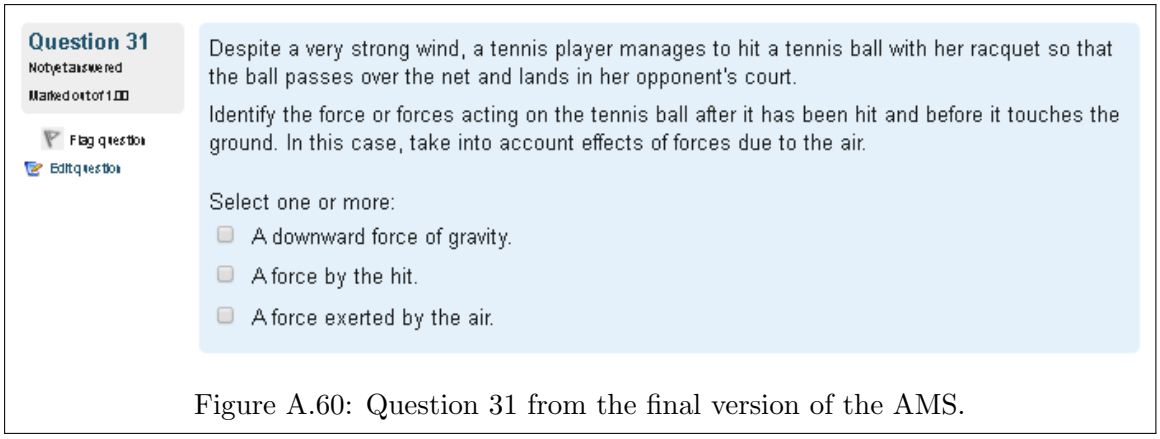


Figure A.60: Question 31 from the final version of the AMS.

13 Appendix B: FCI questions

The FCI questions used in the ECUIP study detailed in **Chapter 3** can be found in this appendix.

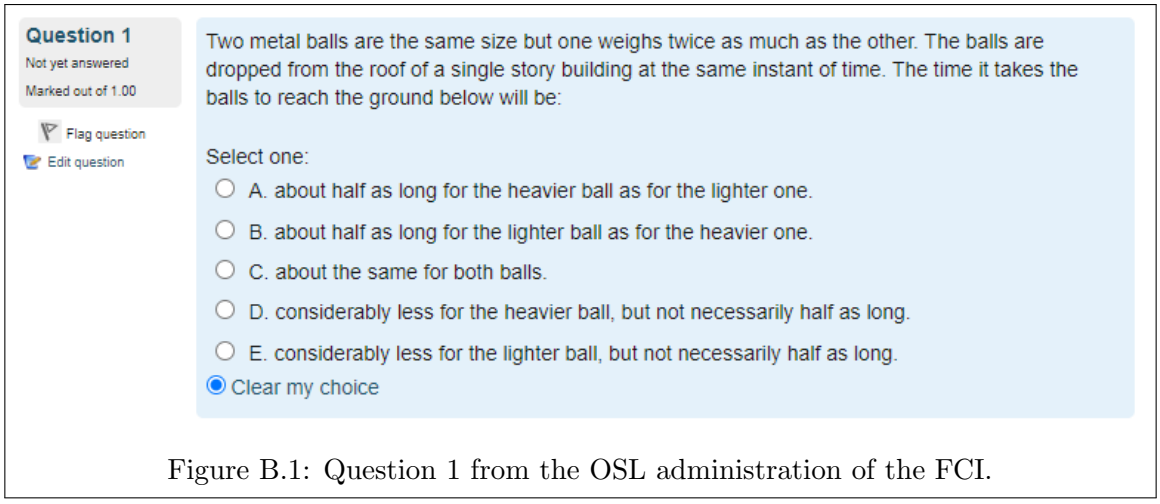


Figure B.1: Question 1 from the OSL administration of the FCI.

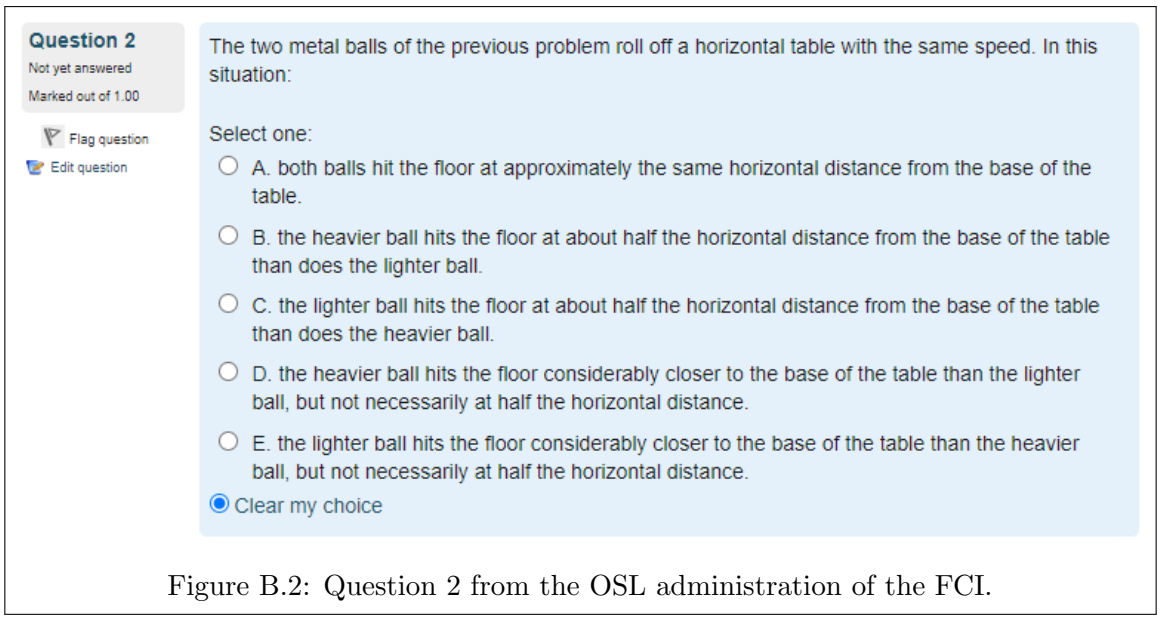


Figure B.2: Question 2 from the OSL administration of the FCI.

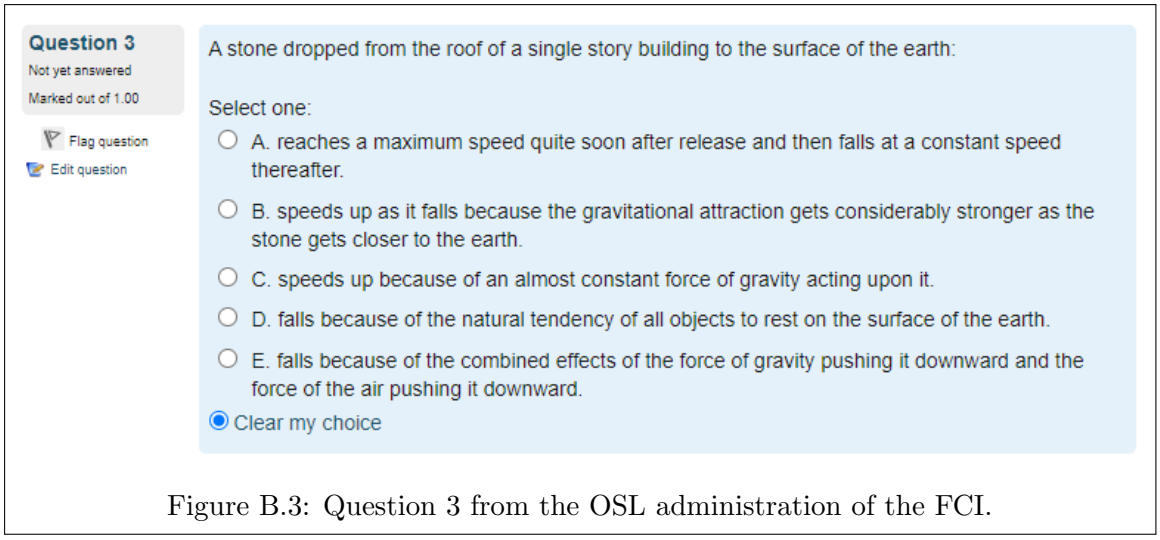


Figure B.3: Question 3 from the OSL administration of the FCI.

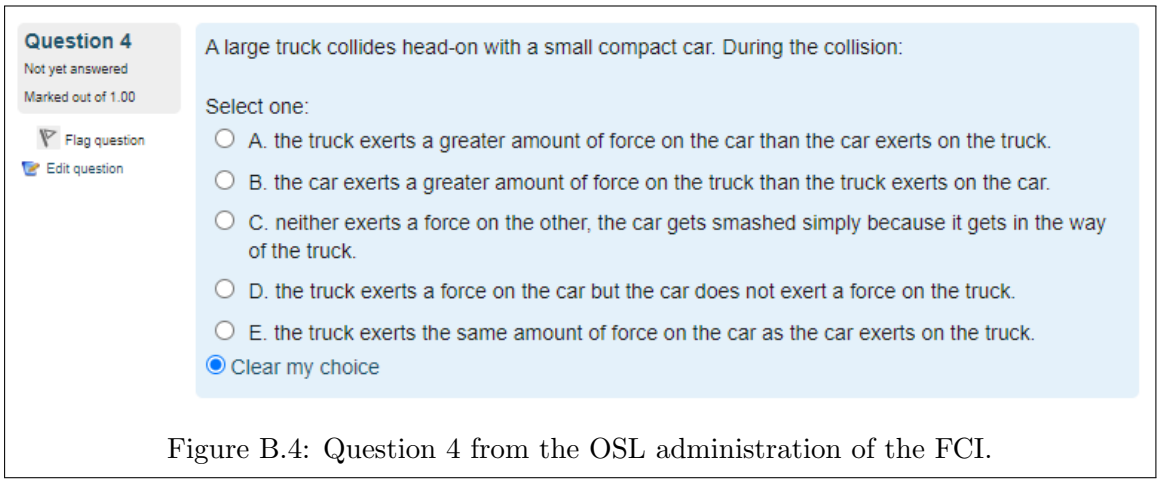


Figure B.4: Question 4 from the OSL administration of the FCI.

Information

Flag question
Edit question

USE THE STATEMENT AND FIGURE BELOW TO ANSWER THE NEXT TWO QUESTIONS (5 and 6).

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with center at *O*.
The channel has been anchored to a frictionless horizontal table top. You are looking down at the table.
Forces exerted by the air are negligible. A ball is shot at high speed into the channel at *P* and exits at *R*.

Consider the following distinct forces:

1. A downward force of gravity
2. A force exerted by the channel pointing from *Q* to *O*.
3. A force in the direction of motion.
4. A force pointing from *O* to *Q*.

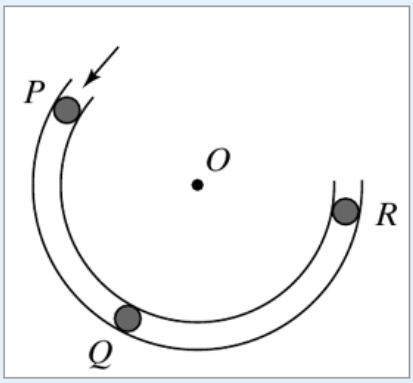


Figure B.5: Prelude information for Question 5 and Question 6 from the OSL administration of the FCI.

Question 5

Not yet answered
Marked out of 1.00

Flag question
Edit question

Which of the above forces is (are) acting on the ball when it is within the frictionless channel at position *Q*?

Select one:

- A. 1 only.
- B. 1 and 2.
- C. 1 and 3.
- D. 1, 2, and 3.
- E. 1, 3, and 4.

Clear my choice

Figure B.6: Question 5 from the OSL administration of the FCI.

Question 6

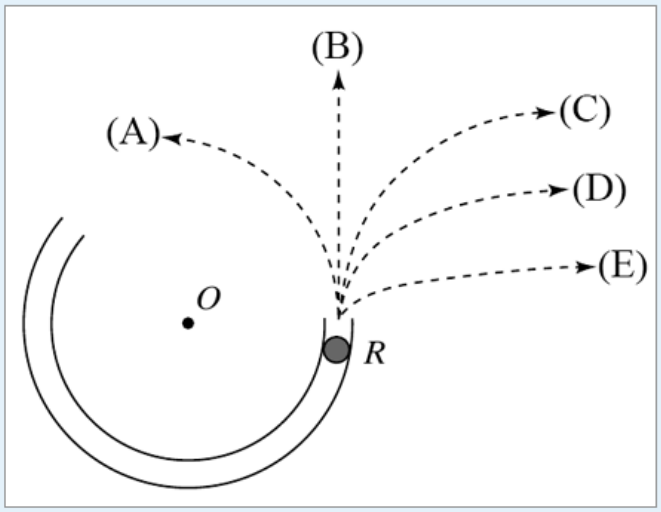
Not yet answered

Marked out of 1.00

Flag question

Edit question

Which path in the accompanying figure would the ball most closely follow after it exits the channel at *R* and moves across the frictionless table top?



Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.

Clear my choice

Figure B.7: Question 6 from the OSL administration of the FCI.

Question 7

Not yet answered

Marked out of 1.00

Flag question

Edit question

A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the accompanying figure.

At the point *P* indicated in the figure, the string suddenly breaks near the ball.

If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

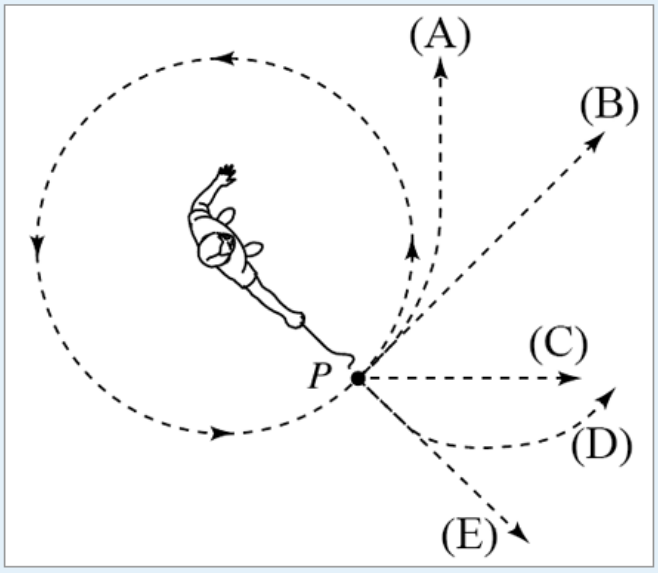


Figure B.8: Question 7 from the OSL administration of the FCI.

Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.
- Clear my choice

Figure B.9: Answer options corresponding to Question 7 from the OSL administration of the FCI.

Information

Flag question

Edit question

USE THE STATEMENT AND FIGURE BELOW TO ANSWER THE NEXT FOUR QUESTIONS (8 through 11).

The figure depicts a hockey puck sliding with constant speed v_o in a straight line from point *P* to point *Q* on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point *Q*, it receives a swift horizontal kick in the direction of the large arrow. Had the puck been at rest at point *Q*, then the kick would have set the puck in horizontal motion with a speed v_k in the direction of the kick.

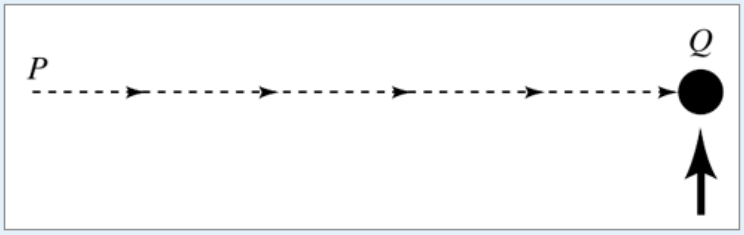


Figure B.10: Prelude information for Question 8, Question 9, Question 10 and Question 11 from the OSL administration of the FCI.

Question 8

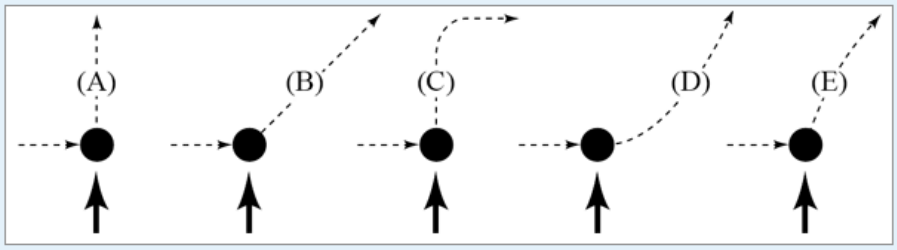
Not yet answered

Marked out of 1.00

Flag question

Edit question

Which of the paths below would the puck most closely follow after receiving the kick?



Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.

Clear my choice

Figure B.11: Question 8 from the OSL administration of the FCI.

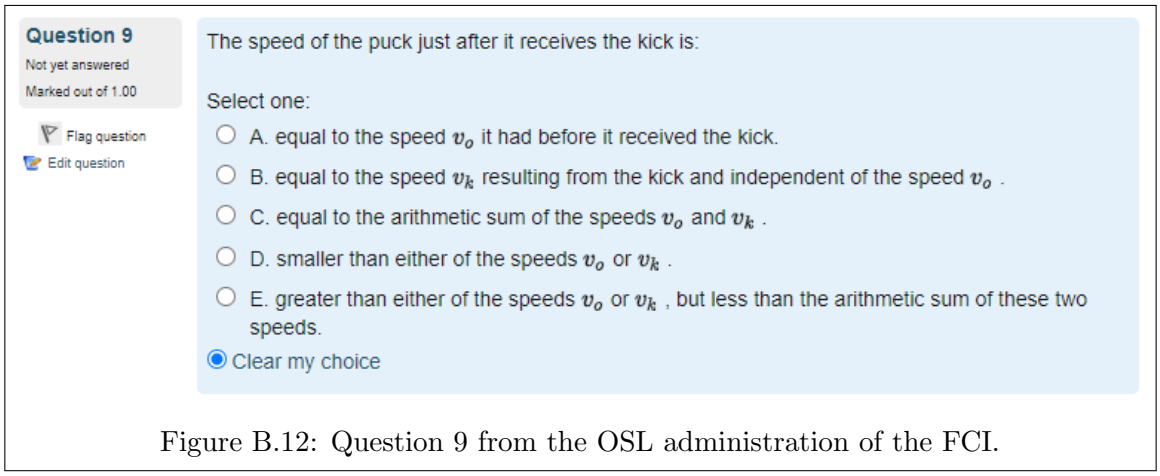


Figure B.12: Question 9 from the OSL administration of the FCI.

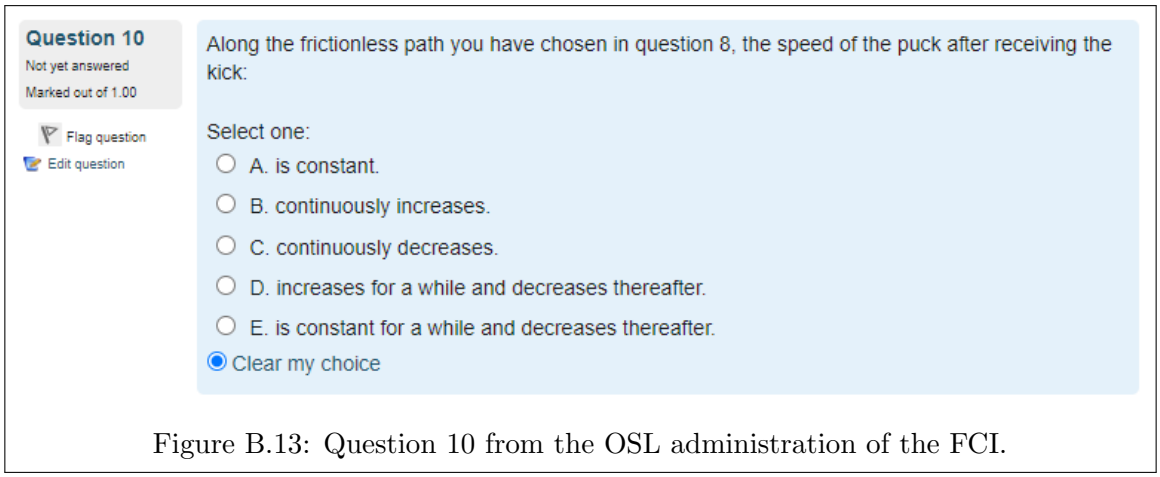


Figure B.13: Question 10 from the OSL administration of the FCI.

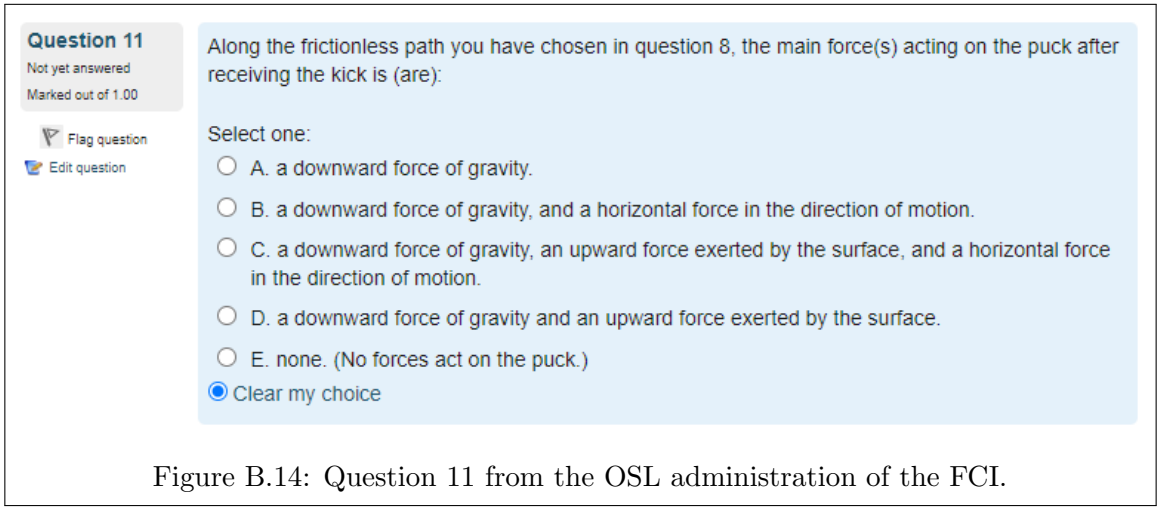


Figure B.14: Question 11 from the OSL administration of the FCI.

Question 12
Not yet answered
Marked out of 1.00

Flag question
Edit question

A ball is fired by a cannon from the top of a cliff as shown in the accompanying figure. Which of the paths would the cannon ball most closely follow?

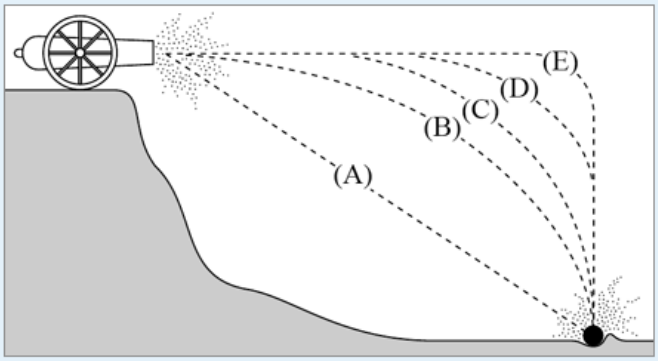


Figure B.15: Question 12 from the OSL administration of the FCI.

Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.
- Clear my choice

Figure B.16: Answer options corresponding to Question 12 from the OSL administration of the FCI.

Question 13
Not yet answered
Marked out of 1.00

Flag question
Edit question

A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are):

Select one:

- A. a downward force of gravity along with a steadily decreasing upward force.
- B. a steadily decreasing upward force from the moment it leaves the boys hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to the earth.
- C. an almost constant downward force of gravity along with an upward force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity.
- D. an almost constant downward force of gravity only.
- E. none of the above. The ball falls back to ground because of its natural tendency to rest on the surface of the earth.
- Clear my choice

Figure B.17: Question 13 from the OSL administration of the FCI.

Question 14

Not yet answered

Marked out of 1.00

Flag question

Edit question

A bowling ball accidentally falls out of the cargo bay of an airliner as it flies along in a horizontal direction.

As observed by a person standing on the ground and viewing the plane as in the the accompanying figure, which path would the bowling ball most closely follow after leaving the airplane?

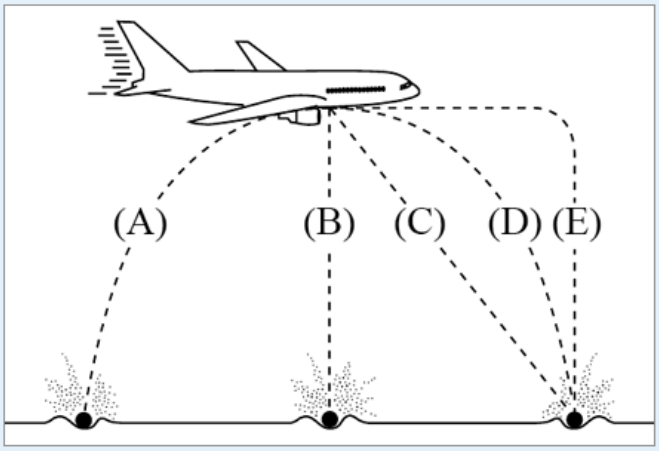


Figure B.18: Question 14 from the OSL administration of the FCI.

Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.
- Clear my choice

Figure B.19: Answer options corresponding to Question 14 from the OSL administration of the FCI.

Information

Flag question
Edit question

USE THE STATEMENT AND FIGURE BELOW TO ANSWER THE NEXT TWO QUESTIONS (15 and 16).

A large truck breaks down out on the road and receives a push back into town by a small compact car as shown in the accompanying figure.

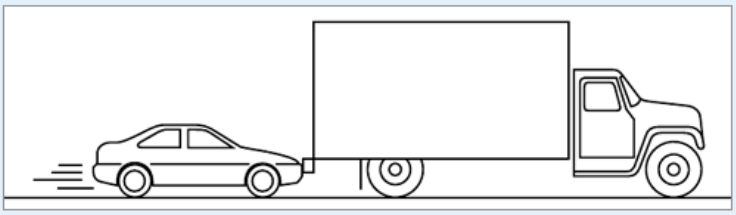


Figure B.20: Prelude information for Question 15 and Question 16 from the OSL administration of the FCI.

Question 15

Not yet answered

Marked out of 1.00

Flag question
Edit question

While the car, still pushing the truck, is speeding up to get up to cruising speed:

Select one:

- A. the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.
- B. the amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.
- C. the amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.
- D. the car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.
- E. neither the car nor the truck exert any force on the other. The truck is pushed forward simply because it is in the way of the car.

Clear my choice

Figure B.21: Question 15 from the OSL administration of the FCI.

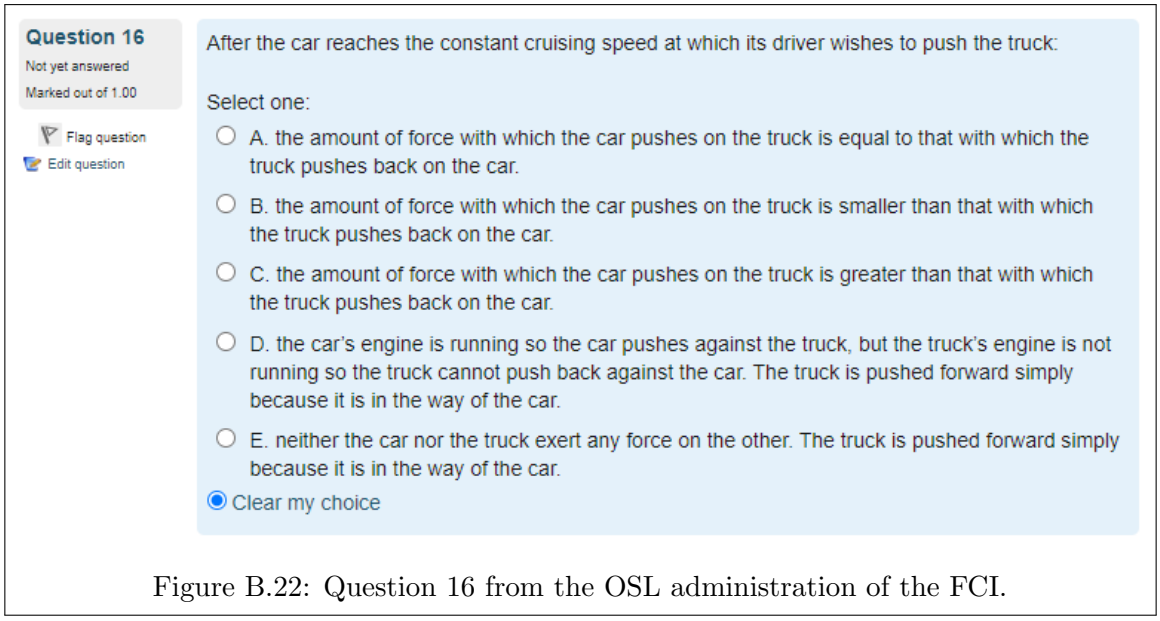


Figure B.22: Question 16 from the OSL administration of the FCI.

Question 17
Not yet answered
Marked out of 1.00

Flag question
Edit question

An elevator is being lifted up an elevator shaft at a constant speed by a steel cable as shown in the accompanying figure. All frictional effects are negligible. In this situation, forces on the elevator are such that:

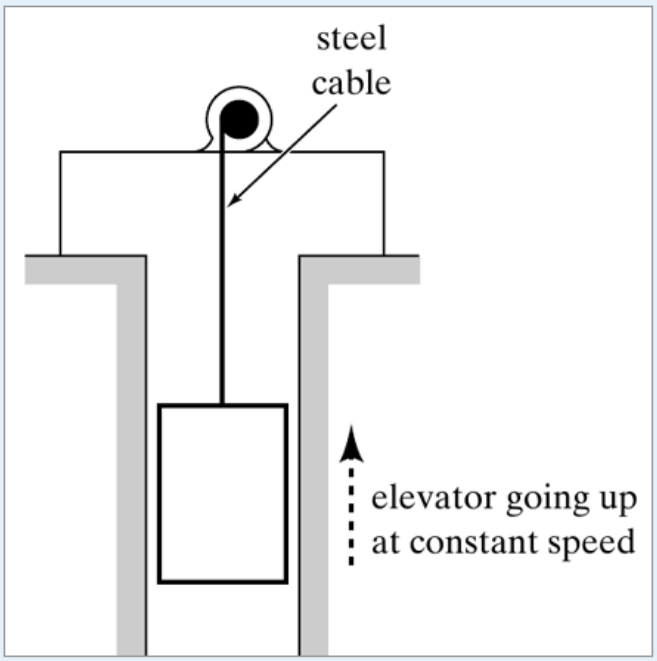


Figure B.23: Question 17 from the OSL administration of the FCI.

Select one:

- A. the upward force by the cable is greater than the downward force of gravity.
- B. the upward force by the cable is equal to the downward force of gravity.
- C. the upward force by the cable is smaller than the downward force of gravity.
- D. the upward force by the cable is greater than the sum of the downward force of gravity and a downward force due to the air.
- E. none of the above. (The elevator goes up because the cable is being shortened, not because an upward force is exerted on the elevator by the cable).

Clear my choice

Figure B.24: Answer options corresponding to Question 17 from the OSL administration of the FCI.

Question 18

Not yet answered

Marked out of 1.00

Flag question

Edit question

The accompanying figure shows a boy swinging on a rope, starting at a point higher than *P*.

Consider the following distinct forces:

1. A downward force of gravity.
2. A force exerted by the rope pointing from *P* to *O*.
3. A force in the direction of the boys motion.
4. A force pointing from *O* to *P*.

Which of the above forces is (are) acting on the boy when he is at position P?

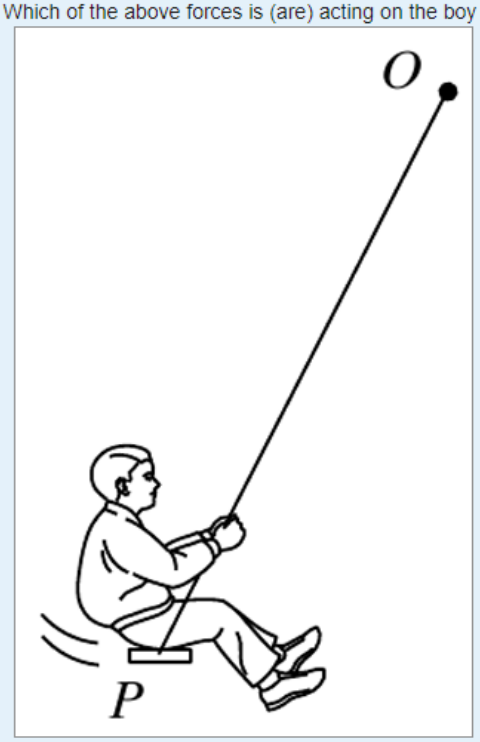


Figure B.25: Question 18 from the OSL administration of the FCI.

Select one:

- A. 1 only.
- B. 1 and 2.
- C. 1 and 3.
- D. 1, 2, and 3.
- E. 1, 3, and 4.

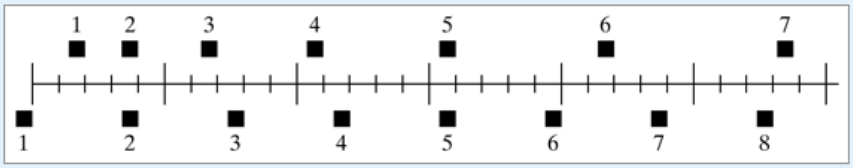
Clear my choice

Figure B.26: Answer options corresponding to Question 18 from the OSL administration of the FCI.

Question 19
Not yet answered
Marked out of 1.00

Flag question
Edit question

The positions of two blocks at successive 0.20-second time intervals are represented by the numbered squares in the accompanying figure. The blocks are moving toward the right.



Do the blocks ever have the same speed?

Select one:

- A. No.
- B. Yes, at instant 2.
- C. Yes, at instant 5.
- D. Yes, at instants 2 and 5.
- E. Yes, at some time during the interval 3 to 4.

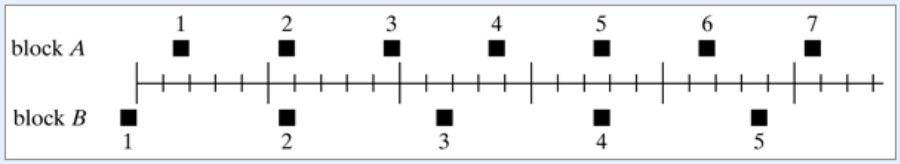
Clear my choice

Figure B.27: Question 19 from the OSL administration of the FCI.

Question 20
Not yet answered
Marked out of 1.00

Flag question
Edit question

The positions of two blocks at successive 0.20-second time intervals are represented by the numbered squares in the accompanying figure. The blocks are moving toward the right.



The accelerations of the blocks are related as follows:

Select one:

- A. The acceleration of *A* is greater than the acceleration of *B*.
- B. The acceleration of *A* equals the acceleration of *B*. Both accelerations are greater than zero.
- C. The acceleration of *B* is greater than the acceleration of *A*.
- D. The acceleration of *A* equals the acceleration of *B*. Both accelerations are zero.
- E. Not enough information is given to answer the question.

Clear my choice

Figure B.28: Question 20 from the OSL administration of the FCI.

Information

Flag question

Edit question

USE THE STATEMENT AND FIGURE BELOW TO ANSWER THE NEXT FOUR QUESTIONS (21 through 24).

A rocket drifts sideways in outer space from point *P* to point *Q* as shown below. The rocket is subject to no outside forces. Starting at position *Q*, the rocket's engine is turned on and produces a constant thrust (force on the rocket) at right angles to the line *PQ*. The constant thrust is maintained until the rocket reaches a point *R* in space.

Figure B.29: Prelude information for Question 21, Question 22, Question 23 and Question 24 from the OSL administration of the FCI.

Question 21

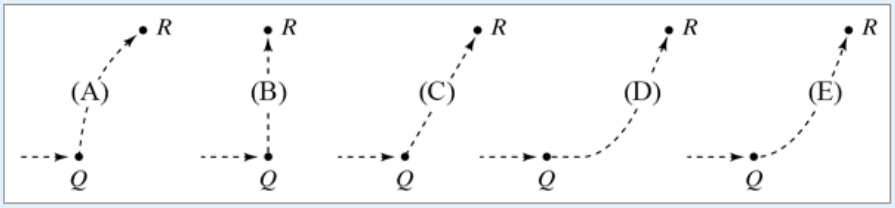
Not yet answered

Marked out of 1.00

Flag question

Edit question

Which of these paths best represents the path of the rocket between points *Q* and *R*?



Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.
- Clear my choice

Figure B.30: Question 21 from the OSL administration of the FCI.

Question 22
Not yet answered
Marked out of 1.00

Flag question
Edit question

As the rocket moves from position *Q* to position *R*, its speed is:

Select one:

- A. constant.
- B. continuously increasing.
- C. continuously decreasing.
- D. increasing for a while and constant thereafter.
- E. constant for a while and decreasing thereafter.

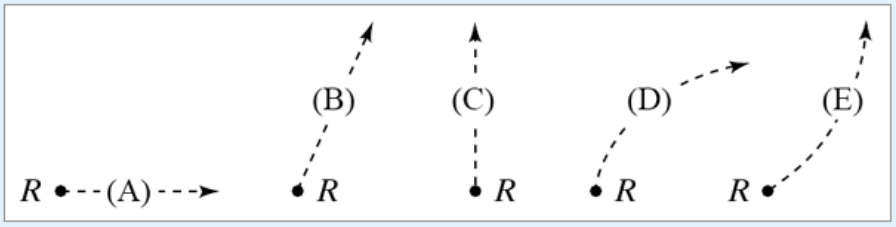
Clear my choice

Figure B.31: Question 22 from the OSL administration of the FCI.

Question 23
Not yet answered
Marked out of 1.00

Flag question
Edit question

At point *R* the rocket's engine is turned off and the thrust immediately drops to zero. Which of these paths will the rocket follow beyond point *R*?



Select one:

- Path A.
- Path B.
- Path C.
- Path D.
- Path E.

Clear my choice

Figure B.32: Question 23 from the OSL administration of the FCI.

Question 24
Not yet answered
Marked out of 1.00

Flag question
Edit question

Beyond position *R* the speed of the rocket is:

Select one:

- A. constant.
- B. continuously increasing.
- C. continuously decreasing.
- D. increasing for a while and constant thereafter.
- E. constant for a while and decreasing thereafter.

Clear my choice

Figure B.33: Question 24 from the OSL administration of the FCI.

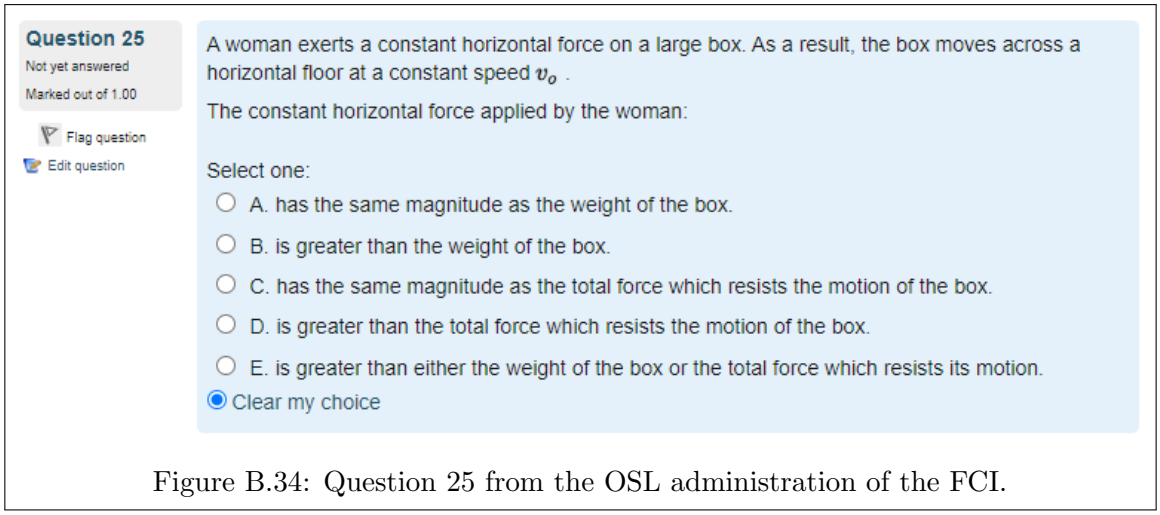


Figure B.34: Question 25 from the OSL administration of the FCI.

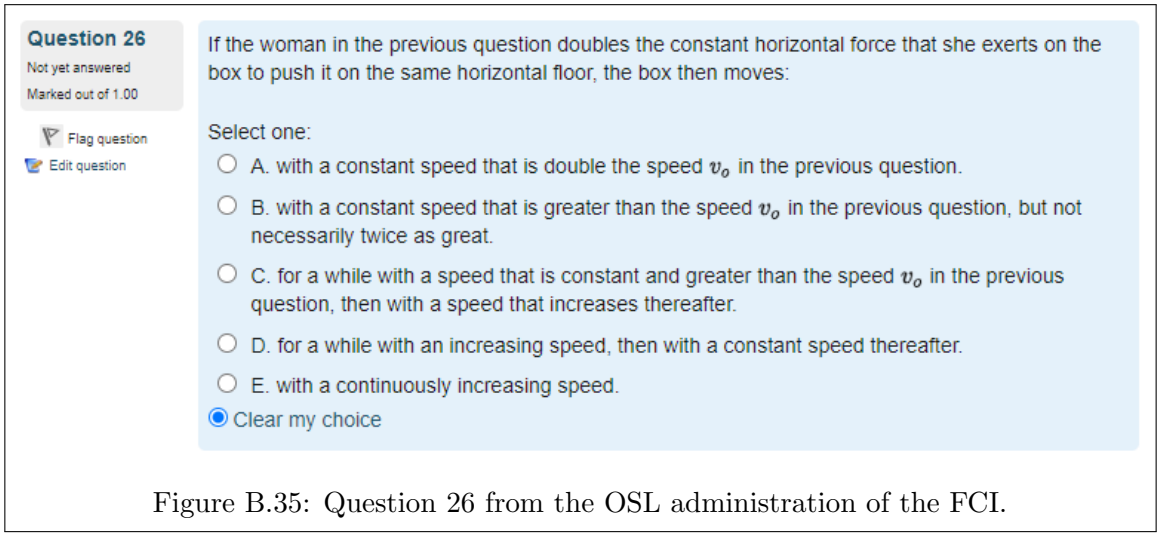


Figure B.35: Question 26 from the OSL administration of the FCI.

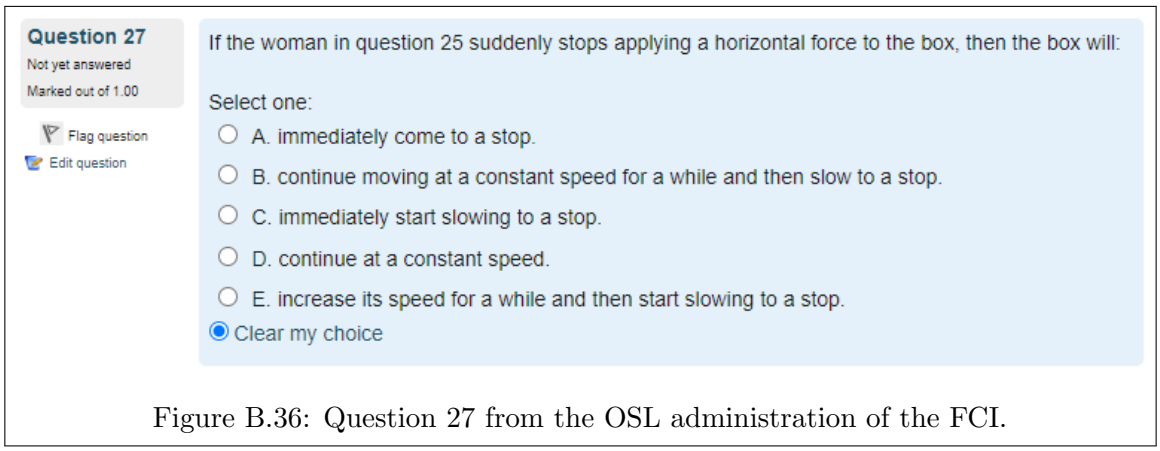


Figure B.36: Question 27 from the OSL administration of the FCI.

Question 28

Not yet answered

Marked out of 1.00

Flag question

Edit question

In the accompanying figure, student *A* has a mass of 95 kg and student *B* has a mass of 77 kg. They sit in identical office chairs facing each other.

Student *A* places his bare feet on the knees of student *B*, as shown. Student *A* then suddenly pushes outward with his feet, causing both chairs to move.

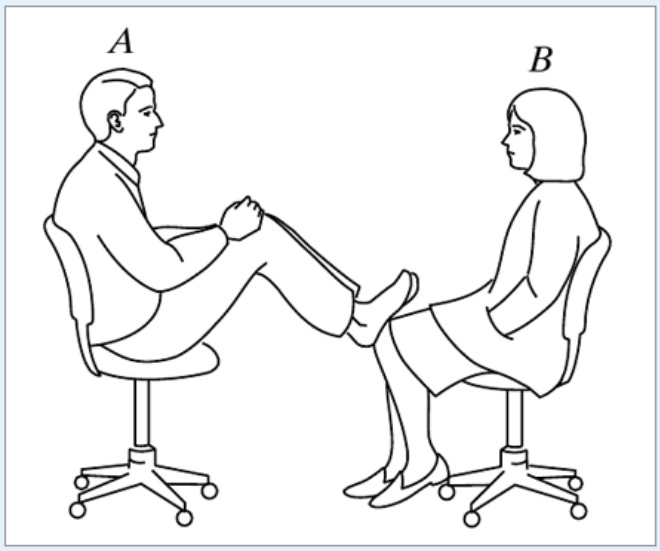


Figure B.37: Question 28 from the OSL administration of the FCI.

During the push and while the students are still touching one another:

Select one:

- A. neither student exerts a force on the other.
- B. student *A* exerts a force on student *B*, but *B* does not exert any force on *A*.
- C. each student exerts a force on the other, but *B* exerts the larger force.
- D. each student exerts a force on the other, but *A* exerts the larger force.
- E. each student exerts the same amount of force on the other.

Clear my choice

Figure B.38: Answer options corresponding to Question 28 from the OSL administration of the FCI.

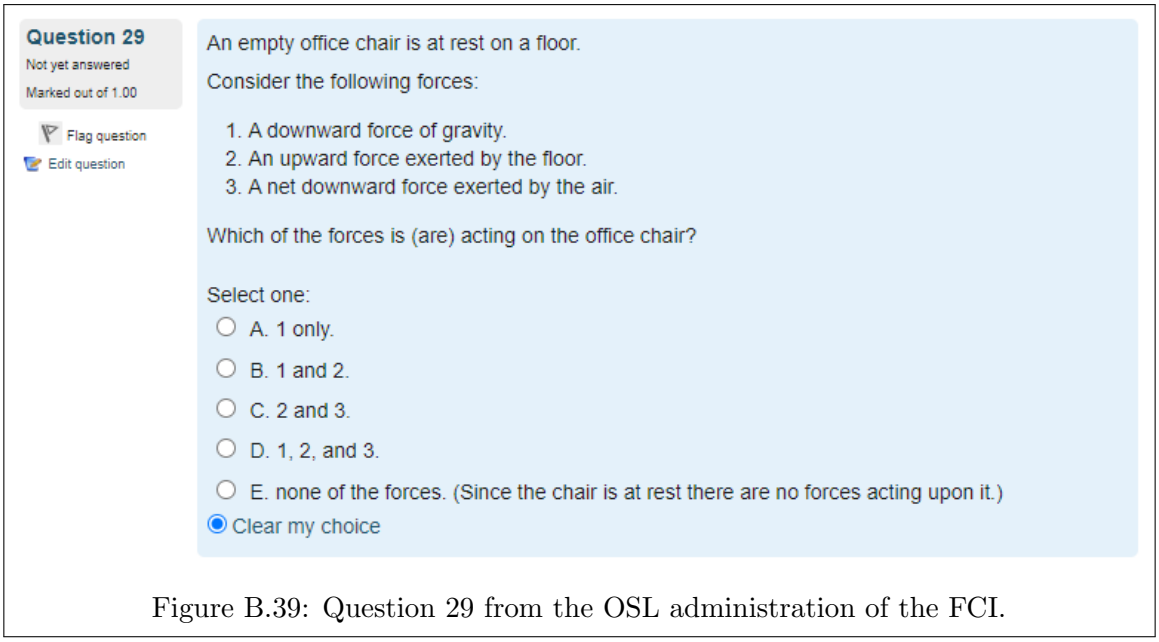


Figure B.39: Question 29 from the OSL administration of the FCI.

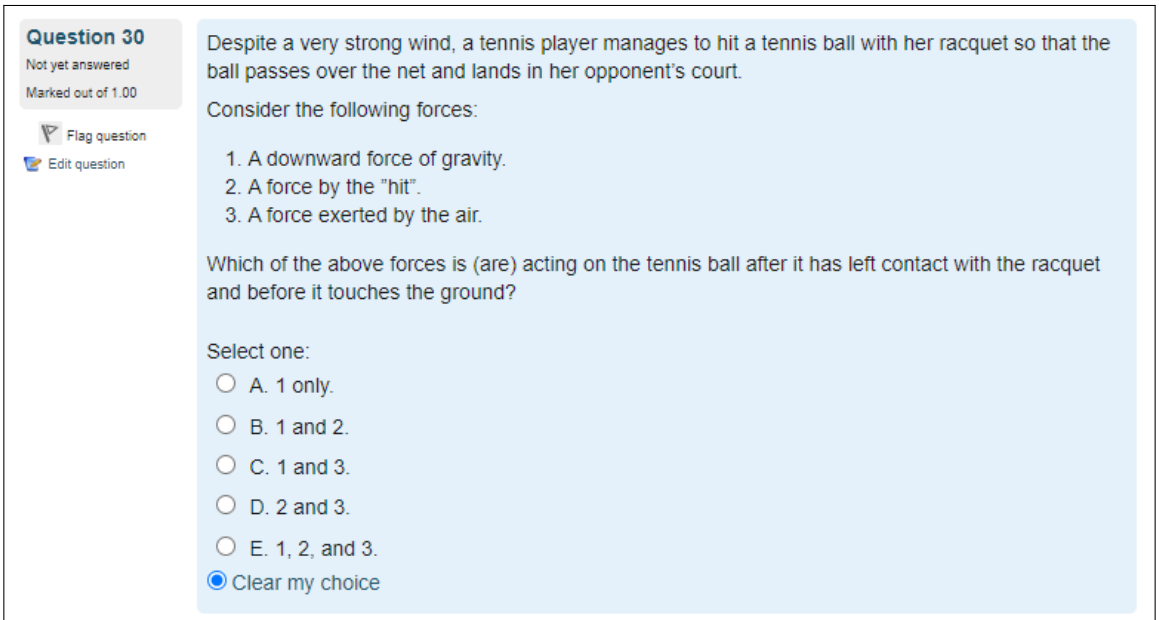


Figure B.40: Question 30 from the OSL administration of the FCI.

14 Appendix C: AMS Version 1 questions

The AMS questions used in the usability testing, CTT and IRR studies detailed in **Chapter 5** and **Chapter 6** can be found in this appendix.

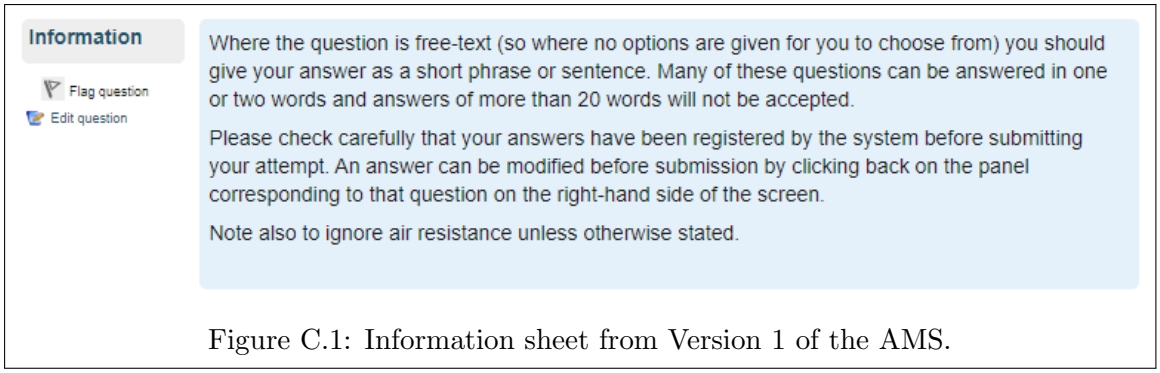


Figure C.1: Information sheet from Version 1 of the AMS.

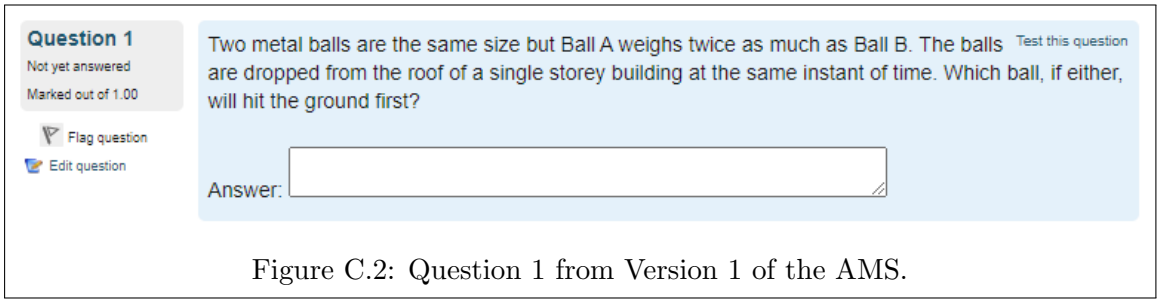


Figure C.2: Question 1 from Version 1 of the AMS.

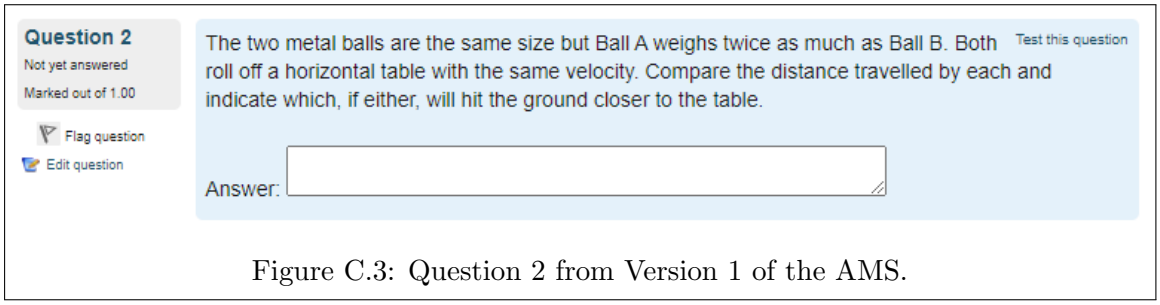


Figure C.3: Question 2 from Version 1 of the AMS.

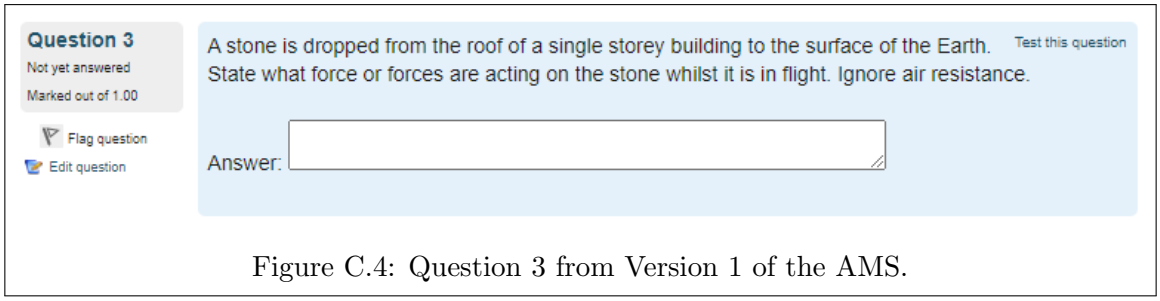


Figure C.4: Question 3 from Version 1 of the AMS.

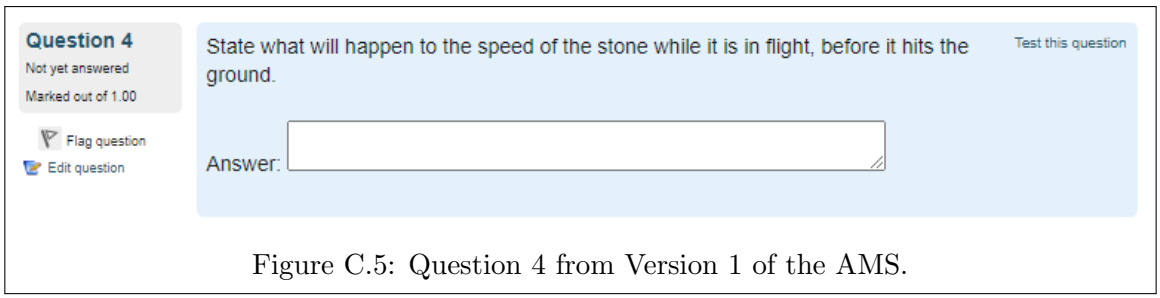


Figure C.5: Question 4 from Version 1 of the AMS.

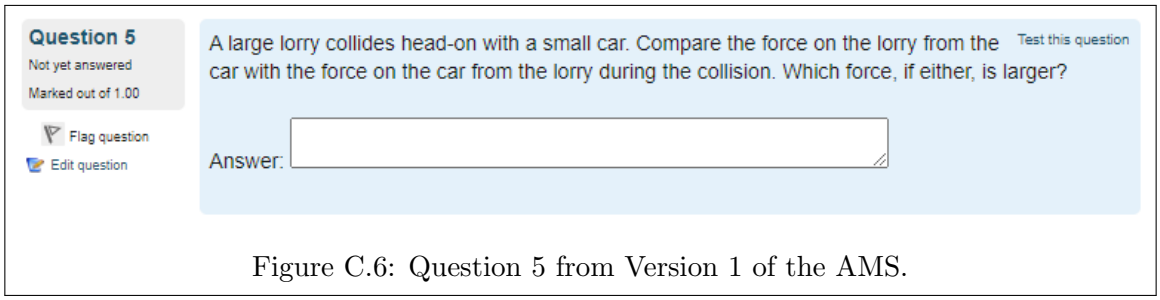


Figure C.6: Question 5 from Version 1 of the AMS.

Question 6

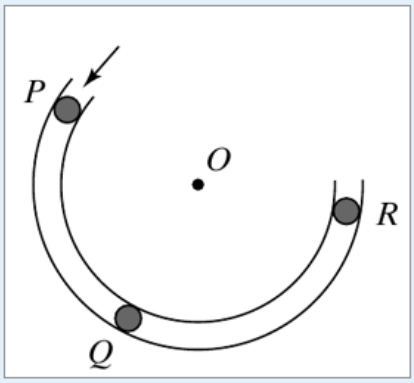
Not yet answered

Marked out of 1.00

Flag question

Edit question

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high speed into the channel at *P* and exits at *R*.



Which of the following forces are acting on the ball when it is in the frictionless channel at point *Q*?

Select one or more:

- ☐ A downward force of gravity.
- ☐ A force pointing from Q to O.
- ☐ A force in the direction of motion.
- ☐ A force pointing from O to Q.
- ☐ An upward force from the table.

Figure C.7: Question 6 from Version 1 of the AMS.

Question 7

Not yet answered

Marked out of 1.00

Flag question

Edit question

Which of the following forces act on the ball just after it emerges from the track at *r*?

Select one or more:

- ☐ a downward force of gravity
- ☐ a force pointing from q to O
- ☐ a force in the direction of motion
- ☐ a force pointing from O to q
- ☐ an upward force from the table

Figure C.8: Question 7 from Version 1 of the AMS.

Question 8

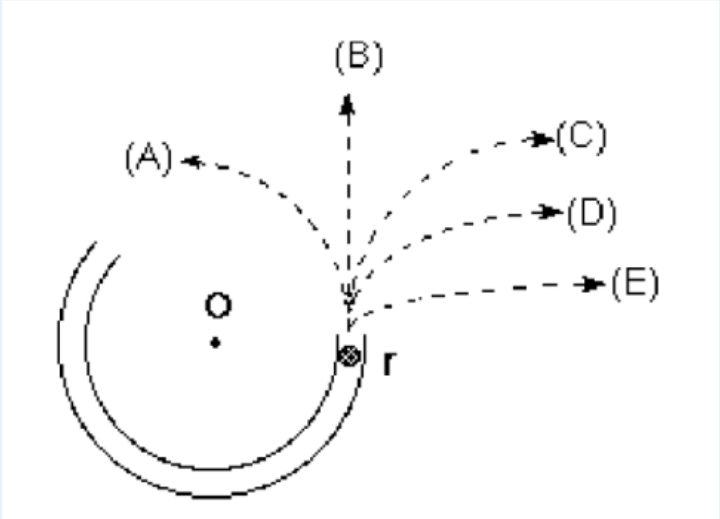
Not yet answered

Marked out of 1.00

Flag question

Edit question

Which path in the figure below would the ball most closely follow after it exits the channel at *r* and moves across the frictionless table top?



Select one:

- A
- B
- C
- D
- E
- Clear my choice

Figure C.9: Question 8 from Version 1 of the AMS.

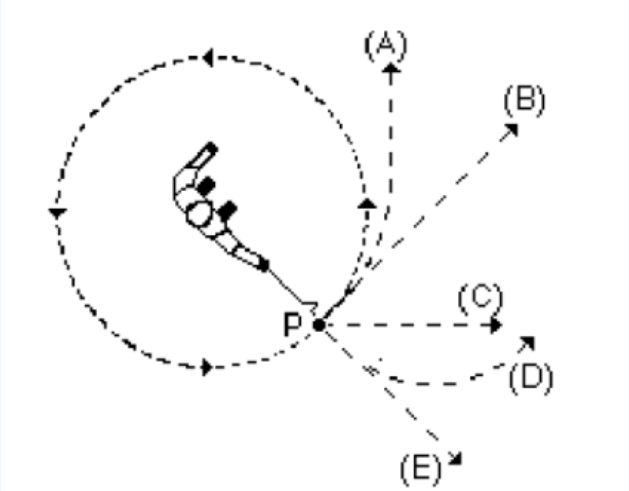
Question 9

Not yet answered

Marked out of 1.00

Flag question

Edit question



A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the above figure. At the point P indicated in the figure, the string suddenly breaks near the ball. If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

Figure C.10: Question 9 from Version 1 of the AMS.

Select one:

- A
- B
- C
- D
- E

Clear my choice

Figure C.11: Answer options corresponding to Question 9 from Version 1 of the AMS.

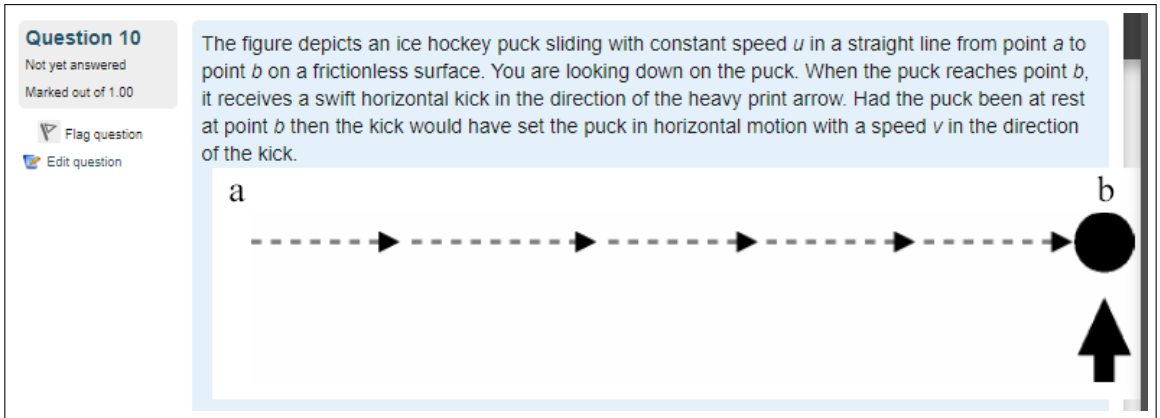


Figure C.12: Prelude information for Question 10, Question 11, Question 12 and Question 13 from Version 1 of the AMS.

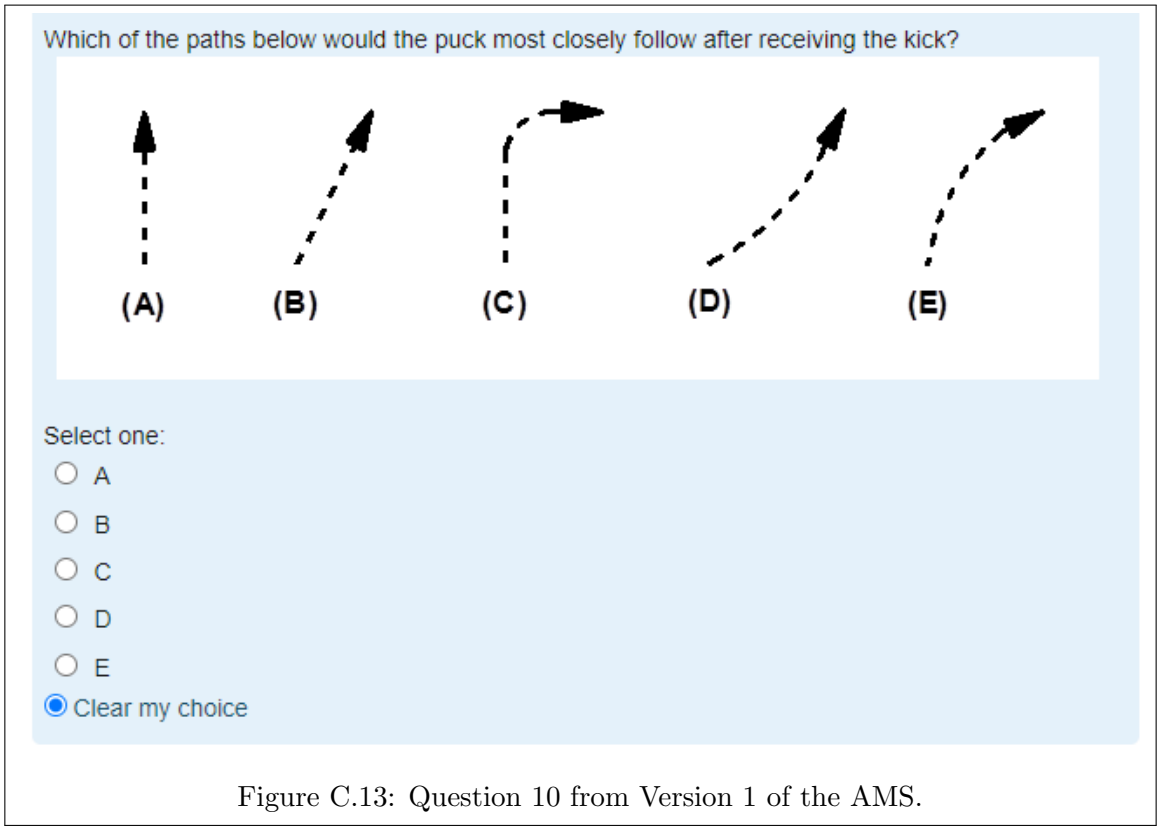


Figure C.13: Question 10 from Version 1 of the AMS.

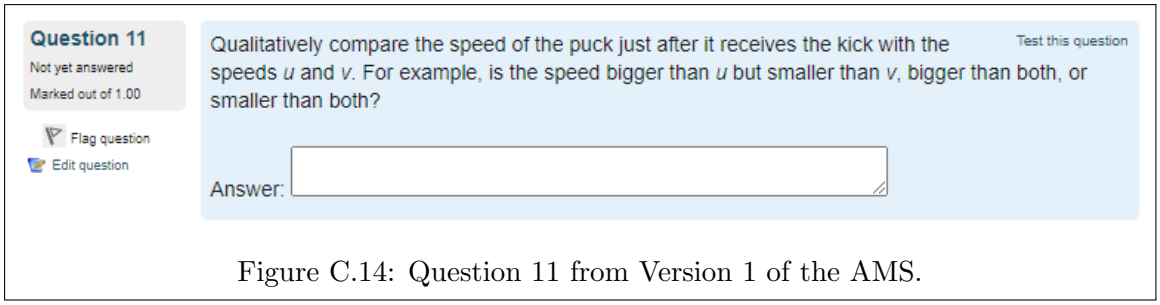


Figure C.14: Question 11 from Version 1 of the AMS.

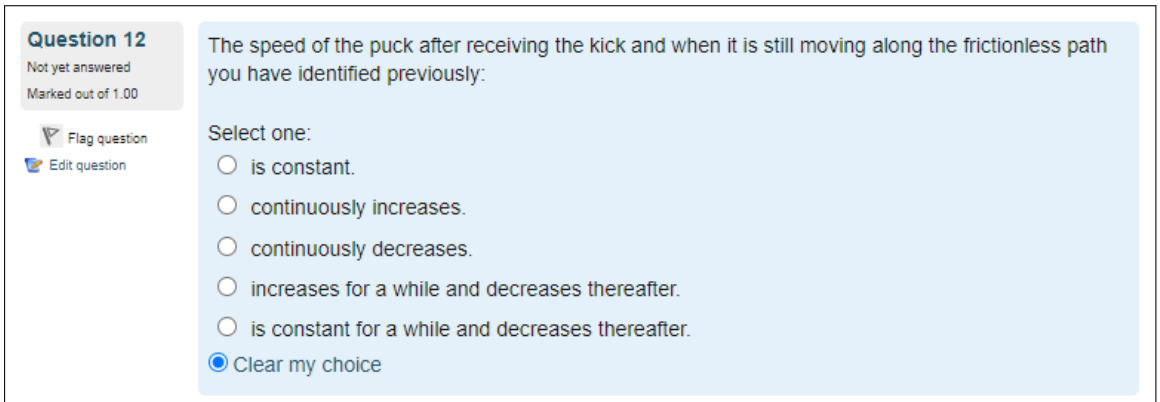


Figure C.15: Question 12 from Version 1 of the AMS.

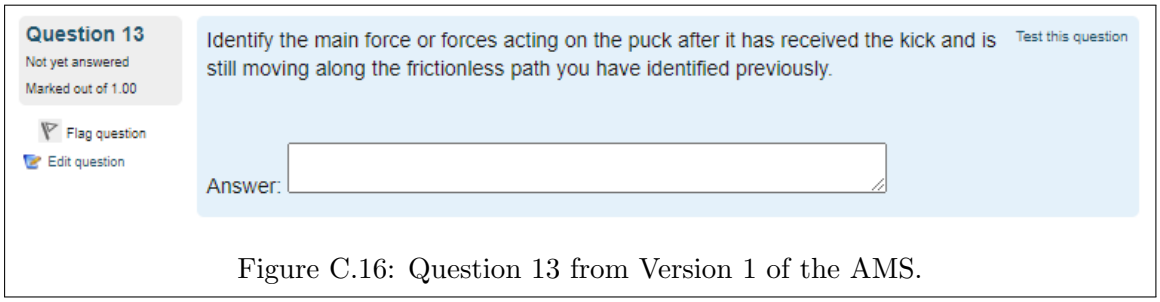


Figure C.16: Question 13 from Version 1 of the AMS.

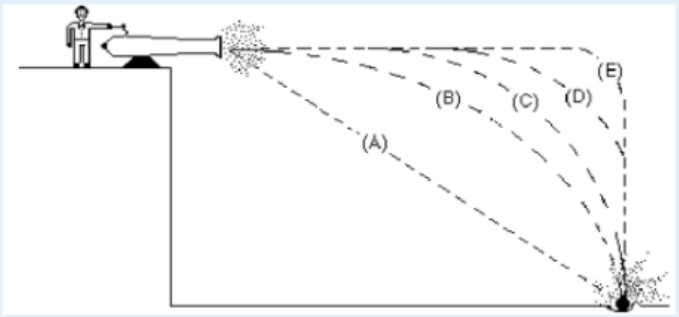
Question 14

Not yet answered

Marked out of 1.00

Flag question

Edit question



A ball is fired by a cannon from the top of a cliff as shown in the figure above. Which of the paths would the cannon ball most closely follow?
Select A, B, C, D or E.

Select one:

- A
- B
- C
- D
- E

Clear my choice

Figure C.17: Question 14 from Version 1 of the AMS.

Question 15

Not yet answered

Marked out of 1.00

Flag question

Edit question

A boy throws a steel ball straight up. What force, or forces, are acting on the ball after it leaves the boy's hand? Ignore air resistance.

Select one:

- A downward force of gravity along with a steadily increasing upward force.
- A steadily decreasing upward force and a steadily increasing downwards force.
- A downward force of gravity along with a steadily decreasing upward force.
- An almost constant downward force of gravity only.
- No forces act on the ball during its motion.

Clear my choice

Figure C.18: Question 15 from Version 1 of the AMS.

Question 16

Not yet answered

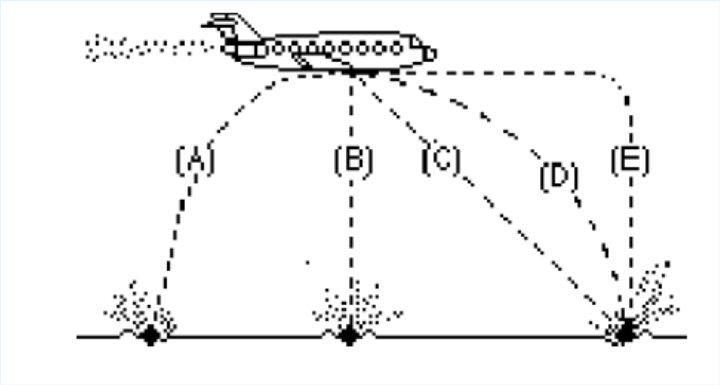
Marked out of 1.00

Flag question

Edit question

A bowling ball accidentally falls out of the cargo bay of an airliner as it flies along in a horizontal direction.

As observed by a person standing on the ground and viewing the plane as in the figure below, which path would the bowling ball most closely follow after leaving the airplane? Select A, B, C, D or E.



Select one:

- A
- B
- C
- D
- E

Clear my choice

Figure C.19: Question 16 from Version 1 of the AMS.

Question 17
Not yet answered
Marked out of 1.00

Flag question
Edit question

A large lorry breaks down and is pushed back into town by a small car, as shown in the figure below. Test this question

Whilst the car, pushing the lorry, is speeding up, how does the force that the car exerts on the lorry compare with the force that the lorry exerts on the car?

Answer:

Figure C.20: Question 17 from Version 1 of the AMS.

Question 18
Not yet answered
Marked out of 1.00

Flag question
Edit question

After the car reaches the constant speed at which its driver wishes to push the lorry:

Select one:

- the amount of force with which the car pushes on the lorry is equal to that with which the lorry pushes back on the car.
- the amount of force with which the car pushes on the lorry is smaller than that with which the lorry pushes back on the car.
- the amount of force with which the car pushes on the lorry is greater than that with which the lorry pushes back on the car.
- the car's engine is running so the car pushes against the lorry, but the lorry's engine is not running so the lorry cannot push back against the car. The lorry is pushed forward simply because it is in the way of the car.
- neither the car nor the lorry exert any force on the other. The lorry is pushed forward simply because it is in the way of the car.

Clear my choice

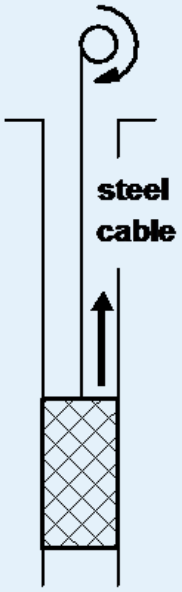
Figure C.21: Question 18 from Version 1 of the AMS.

Question 19
Not yet answered
Marked out of 1.00

Flag question
Edit question

In the diagram below, a lift (shown by the cross hatching) is being hauled up a shaft at a constant speed by a steel cable. All frictional effects are negligible. Identify the force or forces acting on the lift.

Test this question



Answer:

Figure C.22: Question 19 from Version 1 of the AMS.

Question 20
Not yet answered
Marked out of 1.00

Flag question
Edit question

Recall that the lift is going up the shaft at a constant speed. What does this tell you about the forces acting on the lift?

Test this question

Answer:

Figure C.23: Question 20 from Version 1 of the AMS.

Question 21

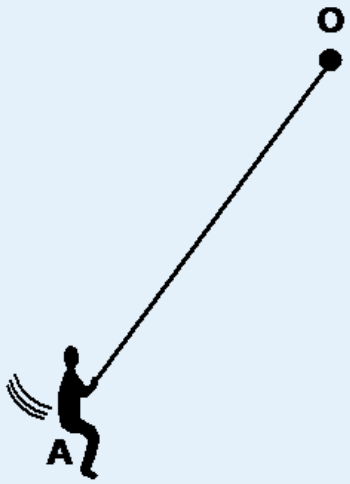
Not yet answered

Marked out of 1.00

Flag question

Edit question

The figure shows a boy swinging on a rope, starting at a point higher than A.



Which of the following distinct forces is (are) acting on the boy when he is at position A?

Select one or more:

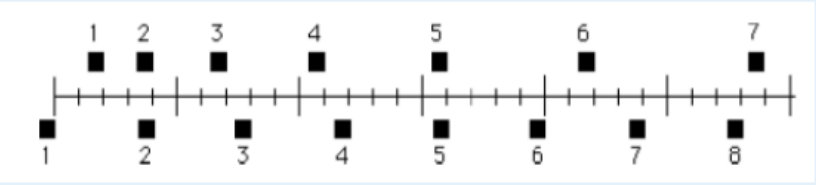
- [ ] A downward force of gravity
- [ ] A force exerted by the rope pointing from A to O
- [ ] A force in the direction of the boy's motion
- [ ] A force pointing from O to A

Figure C.24: Question 21 from Version 1 of the AMS.

Question 22
Not yet answered
Marked out of 1.00

Flag question
Edit question

The positions of two blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Thus, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question



Do the blocks ever have the same speed? If so, describe as accurately as you can when this occurs.

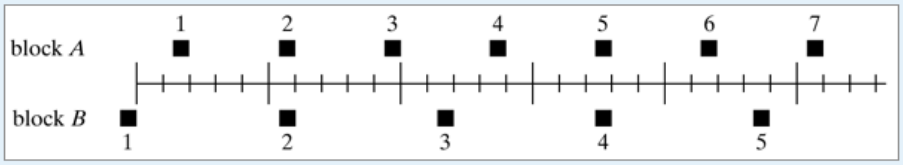
Answer:

Figure C.25: Question 22 from Version 1 of the AMS.

Question 23
Not yet answered
Marked out of 1.00

Flag question
Edit question

Now, the positions of a different pair of blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Again, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question



State if either or both of the blocks are accelerating and if so, which block, if either, has the greater acceleration. Refer to the block above the line as block *A*, and the block below the line as block *B* in your answer.

Answer:

Figure C.26: Question 23 from Version 1 of the AMS.

Question 24

Not yet answered

Marked out of 1.00

Flag question

Edit question

A rocket drifts sideways in outer space from point *a* to point *b* as shown below. The rocket is subject to no outside forces. Starting at position *b*, the rocket's engine is turned on and immediately produces a constant thrust (force on the rocket) at right angles to the line *ab*. The constant thrust is maintained until the rocket reaches a point *c* in space (not shown).

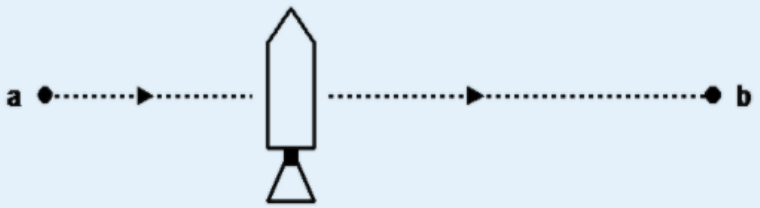
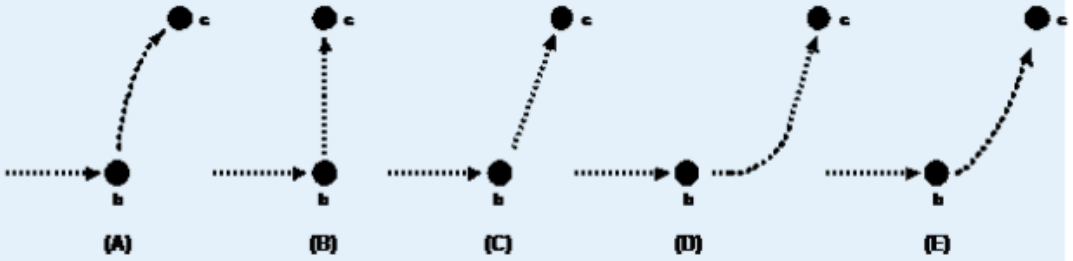


Figure C.27: Prelude information for Question 24, Question 25, Question 26 and Question 27 from Version 1 of the AMS.

Which of the paths below best represents the path of the rocket between points *b* and *c*?

Select A, B, C, D or E.



Select one:

- A
- B
- C
- D
- E
- Clear my choice

Figure C.28: Question 24 from Version 1 of the AMS.

Question 25
Not yet answered
Marked out of 1.00

Flag question
Edit question

As the rocket moves from position *b* to position *c* does its speed increase, decrease or stay the same? Test this question

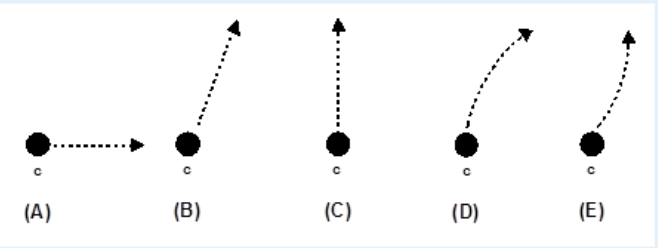
Answer:

Figure C.29: Question 25 from Version 1 of the AMS.

Question 26
Not yet answered
Marked out of 1.00

Flag question
Edit question

At point *c* the rocket's engine is turned off and the thrust immediately drops to zero. Which of the paths below will the rocket follow beyond point *c*? Select A, B, C, D or E.



Select one:
- A
- B
- C
- D
- E
- Clear my choice

Figure C.30: Question 26 from Version 1 of the AMS.

Question 27
Not yet answered
Marked out of 1.00

Flag question
Edit question

What happens to the speed of the rocket beyond position *c*? Test this question

Answer:

Figure C.31: Question 27 from Version 1 of the AMS.

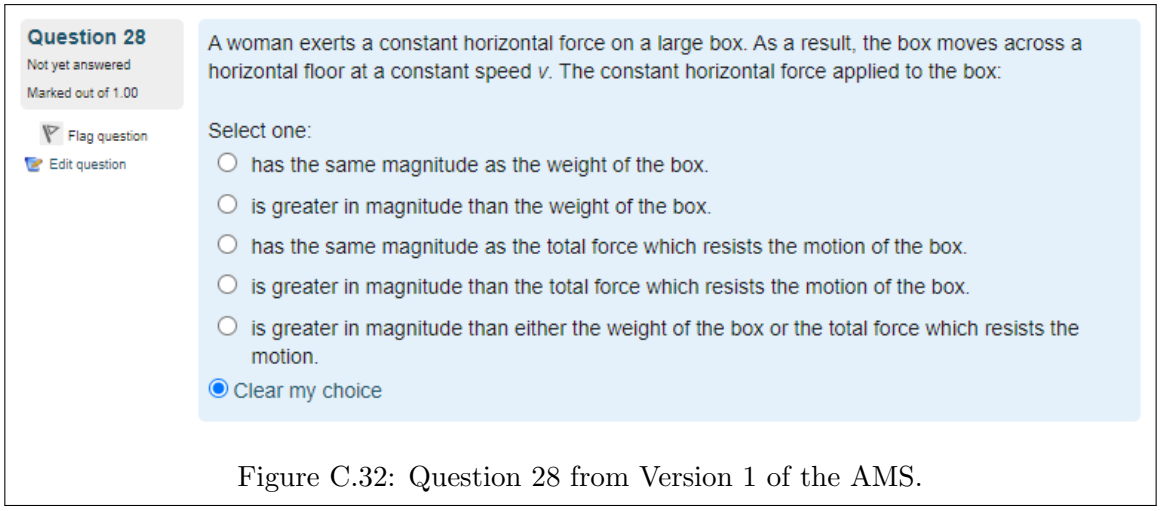


Figure C.32: Question 28 from Version 1 of the AMS.

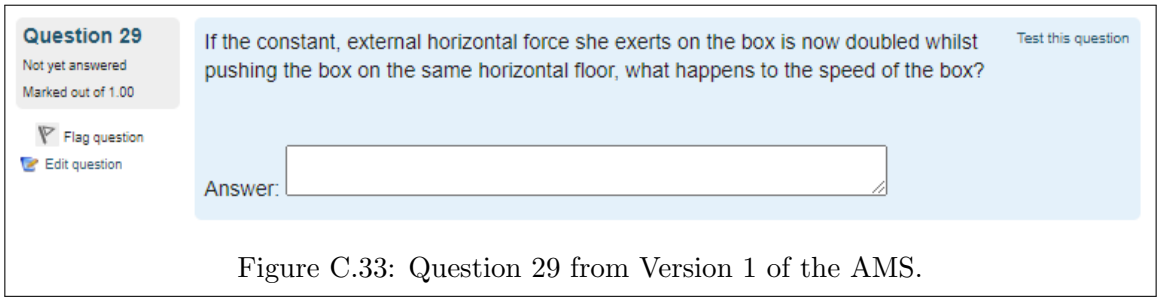


Figure C.33: Question 29 from Version 1 of the AMS.

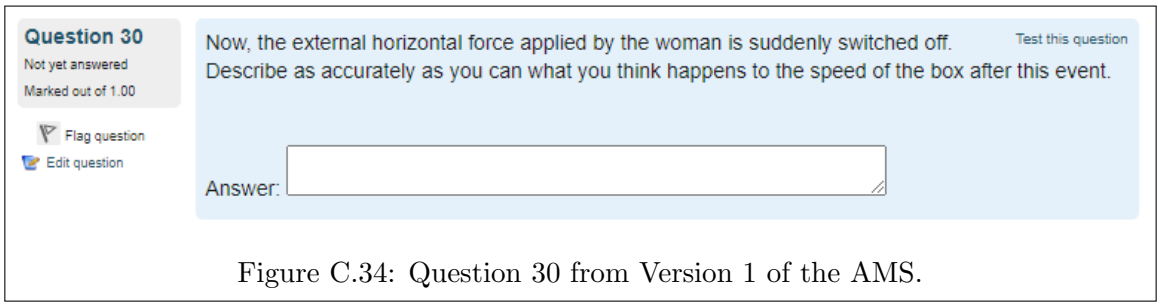


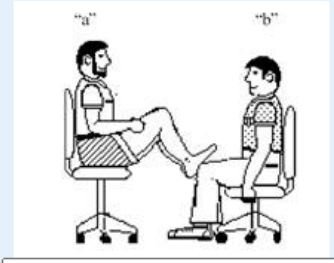
Figure C.34: Question 30 from Version 1 of the AMS.

Question 31
Not yet answered
Marked out of 1.00

Flag question
Edit question

In the figure below, student *a* has a mass of 95 kg and student *b* has a mass of 77 kg. They sit in identical office chairs facing each other. Student *a* places his bare feet on the knees of student *b*, as shown. Student *a* then suddenly pushes outward with his feet, causing both chairs to move. What can you say about the amount of force each student exerts on the other during the push?

Test this question



Answer:

Figure C.35: Question 31 from Version 1 of the AMS.

Question 32
Not yet answered
Marked out of 1.00

Flag question
Edit question

An empty office chair is at rest on a floor. What force or forces act on the office chair?

Test this question

Answer:

Figure C.36: Question 32 from Version 1 of the AMS.

Question 33
Not yet answered
Marked out of 1.00

Flag question
Edit question

Despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court.

Identify the force or forces acting on the tennis ball after it has been hit and before it touches the ground. In this case, take into account effects of forces due to the air.

Select one or more:

- [ ] A downward force of gravity.
- [ ] A force by the hit.
- [ ] A force exerted by the air.

Figure C.37: Question 33 from Version 1 of the AMS.

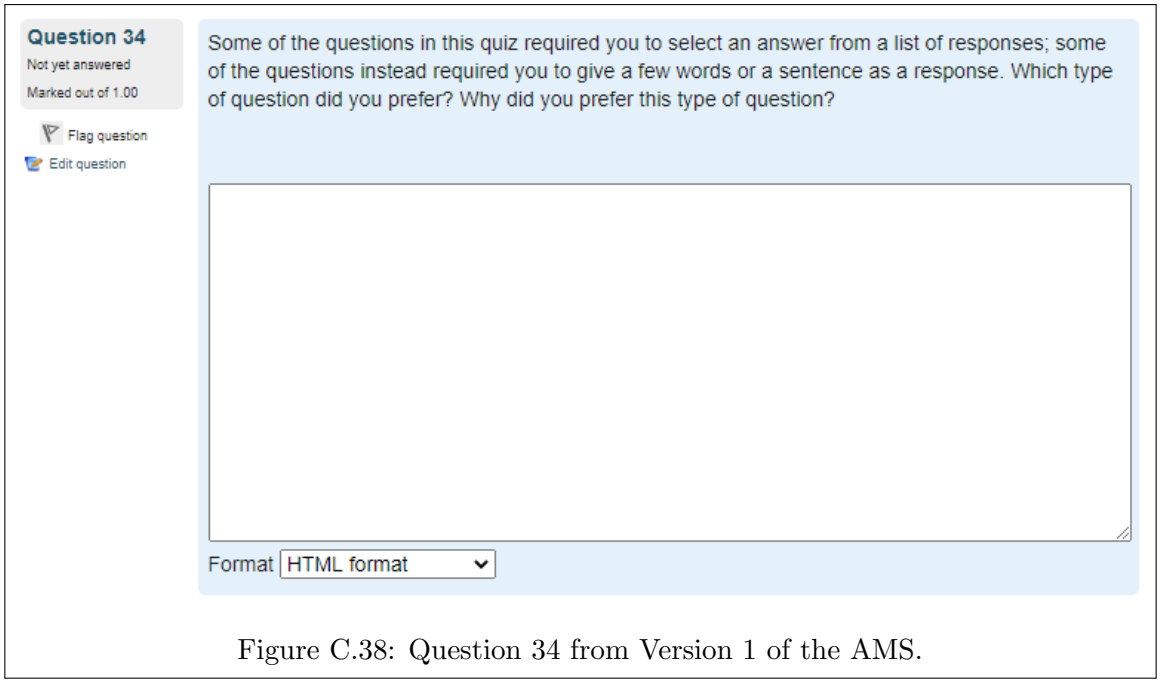


Figure C.38: Question 34 from Version 1 of the AMS.

15 Appendix D: AMS Version 2 questions

The AMS questions used in the CTT and IRR studies detailed in **Chapter 7** can be found in this appendix. Note that the *standardized AMS question numbering* is used here, meaning that there is no Q6, since this question was not present in Version 2 of the AMS.

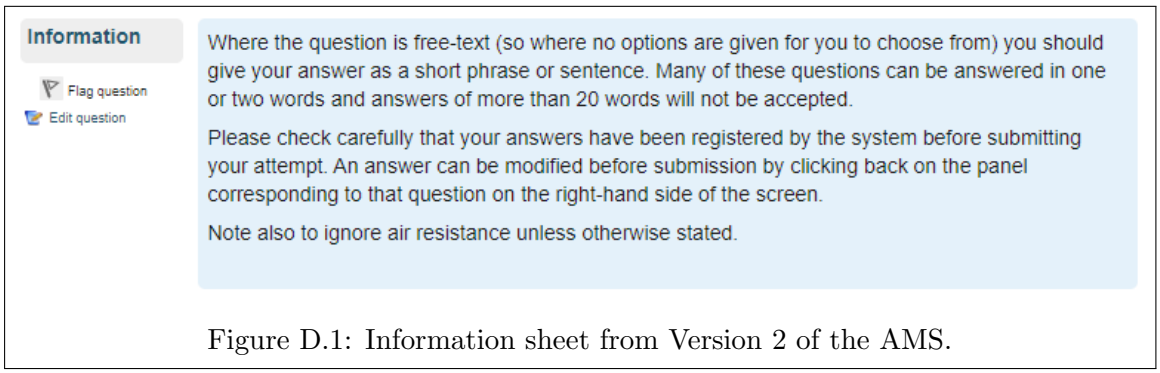


Figure D.1: Information sheet from Version 2 of the AMS.

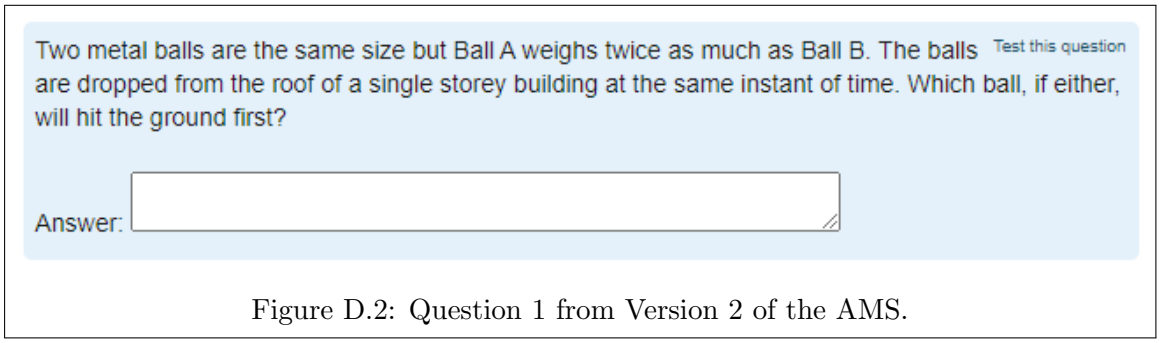


Figure D.2: Question 1 from Version 2 of the AMS.

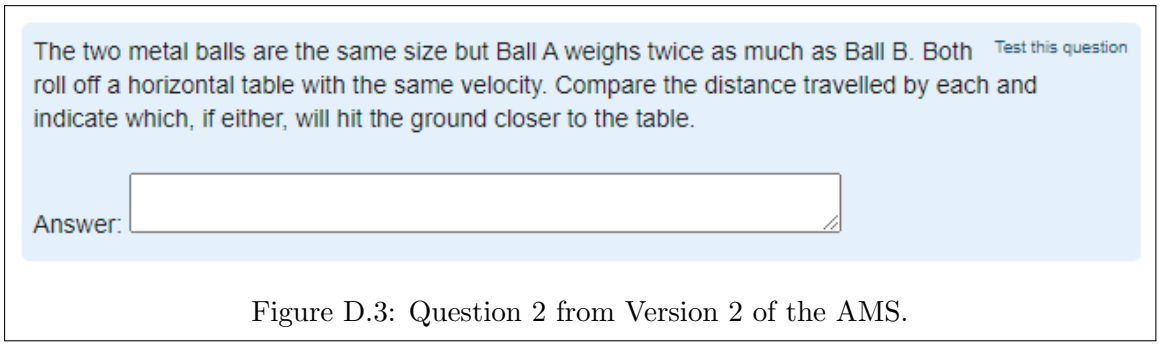


Figure D.3: Question 2 from Version 2 of the AMS.

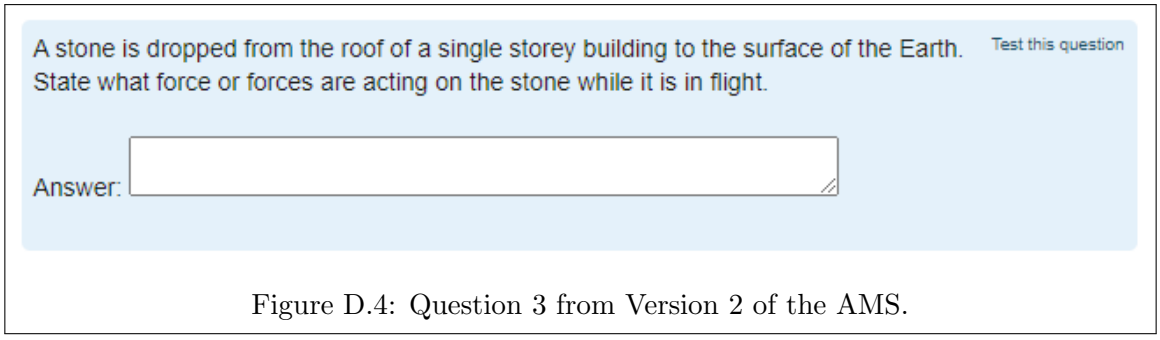


Figure D.4: Question 3 from Version 2 of the AMS.

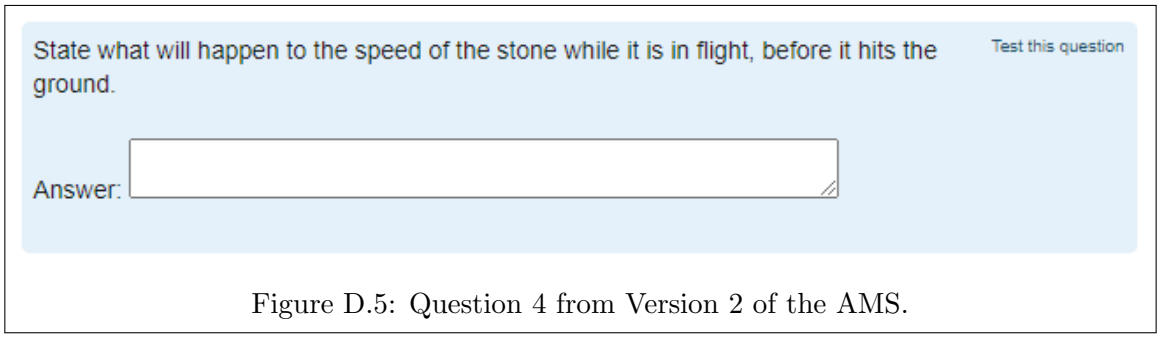


Figure D.5: Question 4 from Version 2 of the AMS.

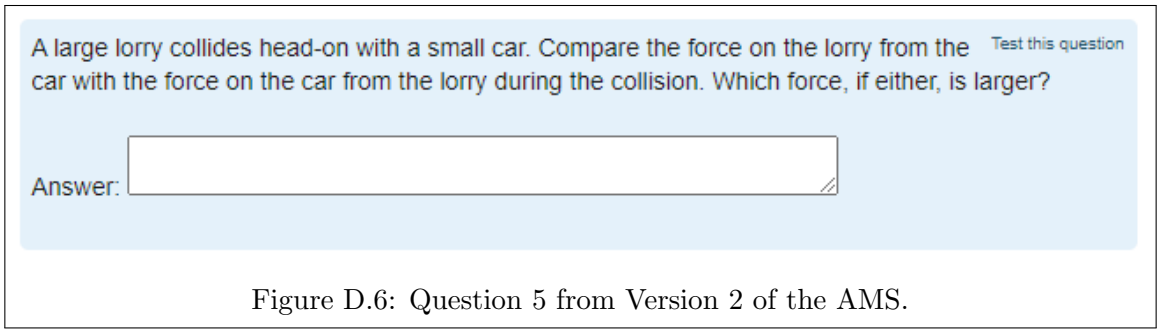
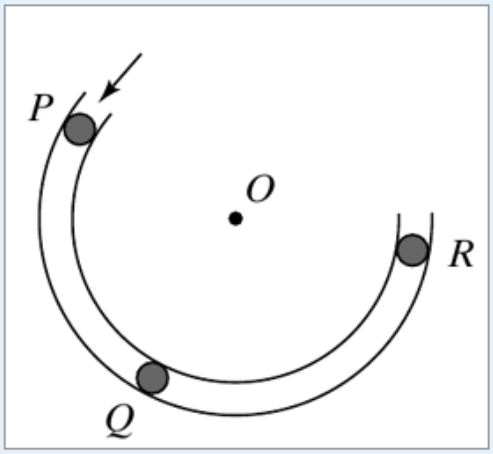


Figure D.6: Question 5 from Version 2 of the AMS.

The accompanying figure shows a frictionless channel in the shape of a segment of a circle
with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high
speed into the channel at *P* and exits at *R*.

Test this question



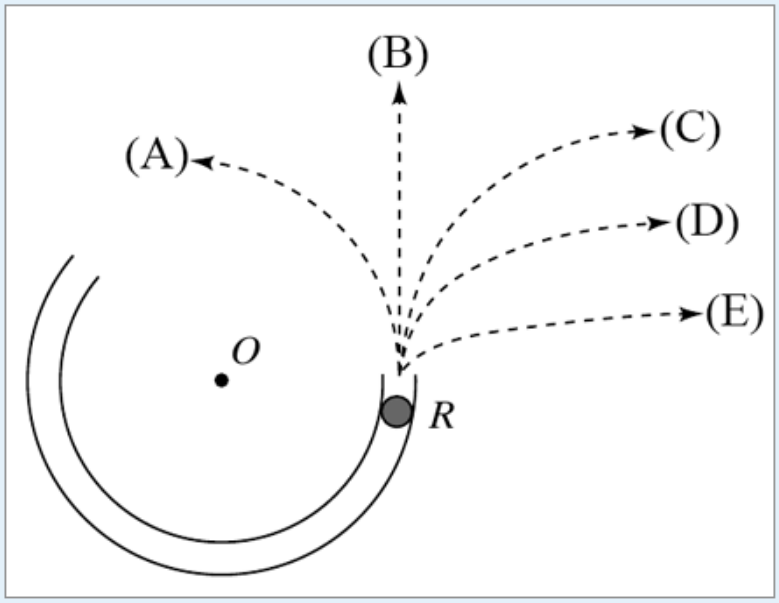
Identify the force or forces acting on the ball after it emerges from the track at R.

Answer:

Figure D.7: Question 7 from Version 2 of the AMS.

Which path in the figure below would the ball most closely follow after it exits the channel at *R* and moves across the frictionless table top?

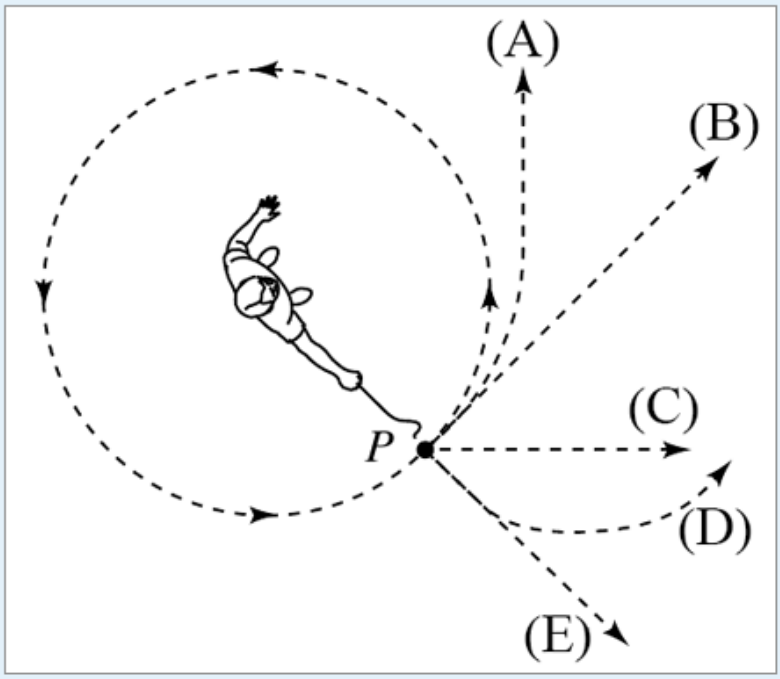
Test this question



Answer:

Figure D.8: Question 8 from Version 2 of the AMS.

Test this question



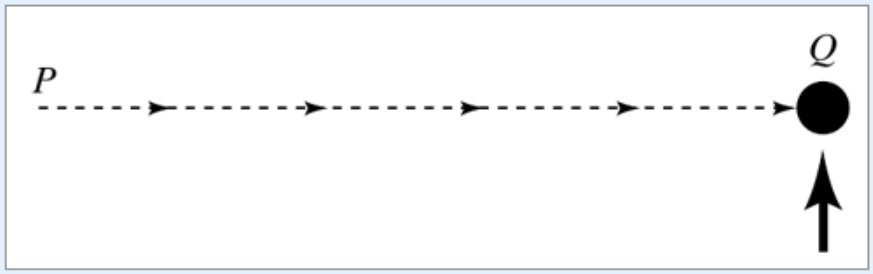
A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the above figure. At the point P indicated in the figure, the string suddenly breaks near the ball. If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

Answer:

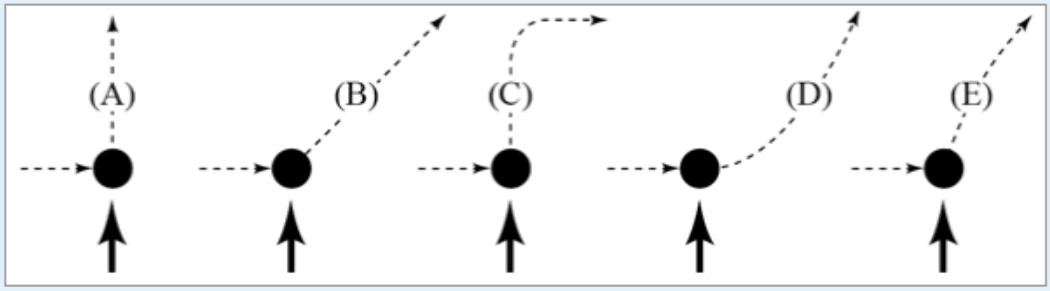
Figure D.9: Question 9 from Version 2 of the AMS.

The figure depicts an ice hockey puck sliding with constant speed u in a straight line from point P to point Q on a frictionless surface. You are looking down on the puck. When the puck reaches point Q, it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point Q then the kick would have set the puck in horizontal motion with a speed w in the direction of the kick.

Test this question



Which of the paths below would the puck most closely follow after receiving the kick?



Answer:

Figure D.10: Question 10 from Version 2 of the AMS.

Qualitatively compare the speed of the puck just after it receives the kick with the speeds u and v. For example, is the speed bigger than u but smaller than v, bigger than both, or smaller than both?

Test this question

Answer:

Figure D.11: Question 11 from Version 2 of the AMS.

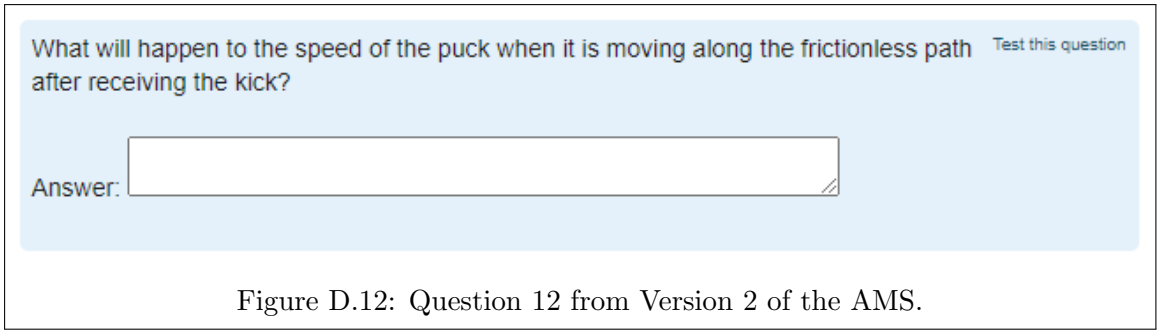


Figure D.12: Question 12 from Version 2 of the AMS.

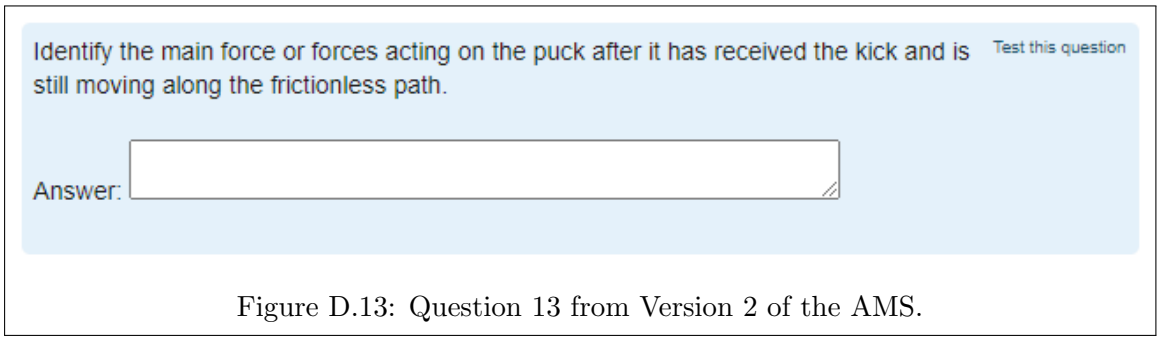
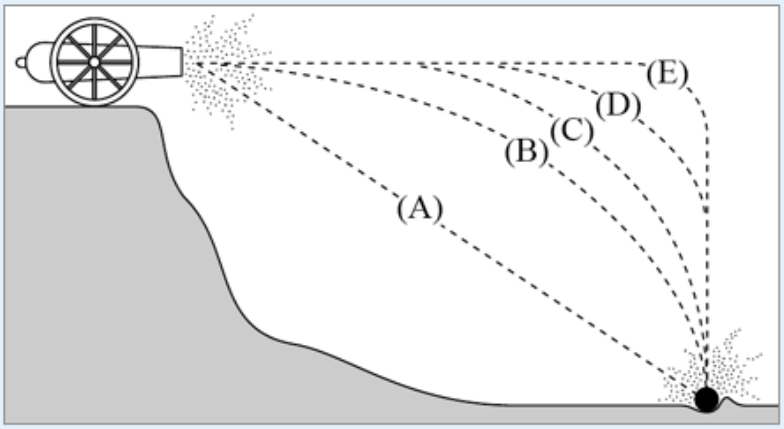


Figure D.13: Question 13 from Version 2 of the AMS.

Test this question



A ball is fired by a cannon from the top of a cliff as shown in the figure above. Which of the paths would the cannon ball most closely follow?

Answer:

Figure D.14: Question 14 from Version 2 of the AMS.

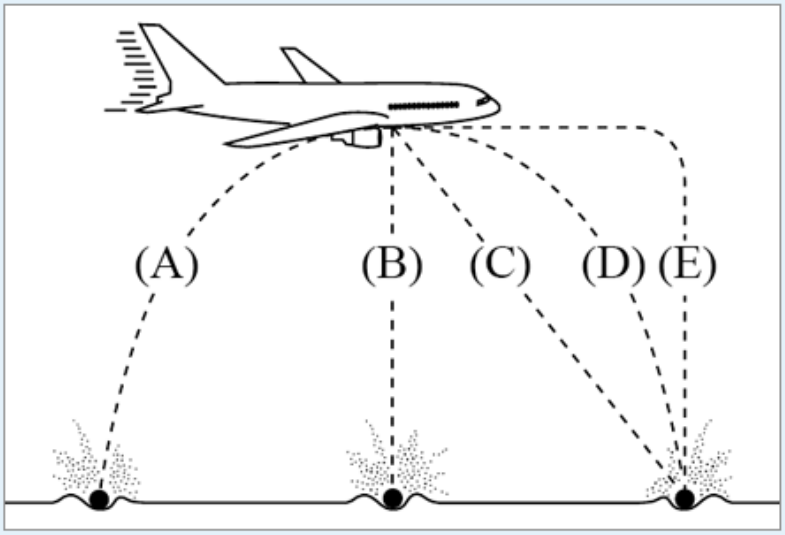
A boy throws a steel ball straight up. What force, or forces, are acting on the ball after it leaves the boy's hand? Test this question

Answer:

Figure D.15: Question 15 from Version 2 of the AMS.

A bowling ball accidentally falls out of the cargo bay of an airliner as it flies along in a horizontal direction. Test this question

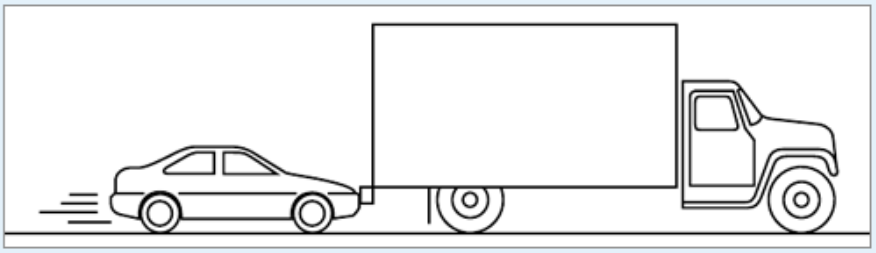
As observed by a person standing on the ground and viewing the plane as in the figure below, which path would the bowling ball most closely follow after leaving the airplane?



Answer:

Figure D.16: Question 16 from Version 2 of the AMS.

A large lorry breaks down and is pushed back into town by a small car, as shown in the figure below. Test this question



While the car, pushing the lorry, is speeding up, how does the force that the car exerts on the lorry compare with the force that the lorry exerts on the car?

Answer:

Figure D.17: Question 17 from Version 2 of the AMS.

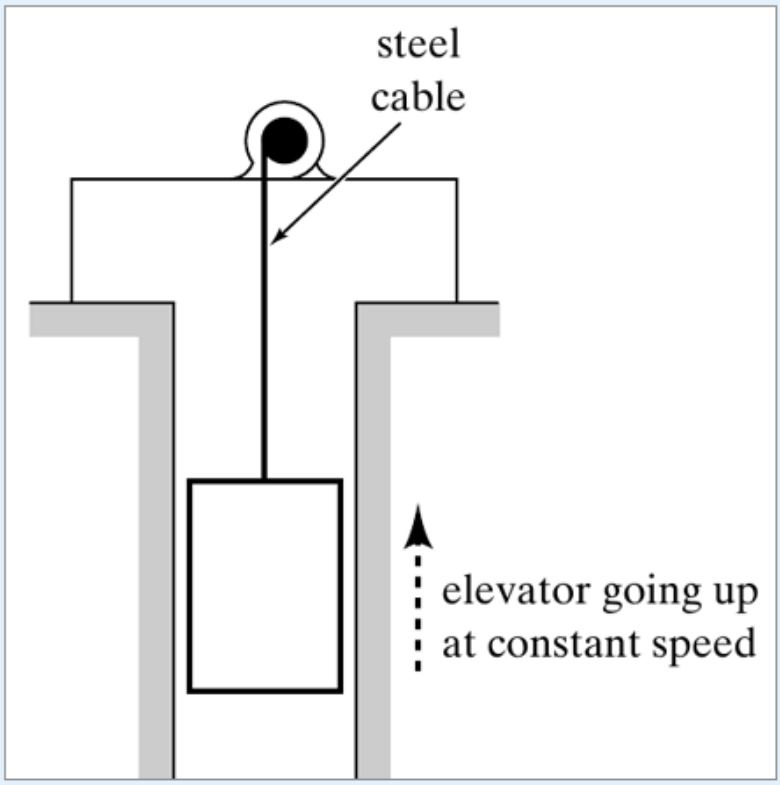
After the car reaches the constant speed at which its driver wishes to push the lorry, how does the force that the car exerts on the lorry compare with the force that the lorry exerts on the car? Test this question

Answer:

Figure D.18: Question 18 from Version 2 of the AMS.

In the diagram below, a lift is being hauled up a shaft at a constant speed by a steel cable. Identify the force or forces acting on the lift.

Test this question



Answer:

Figure D.19: Question 19 from Version 2 of the AMS.

Recall that the lift is going up the shaft at a constant speed. What does this tell you about the forces acting on the lift?

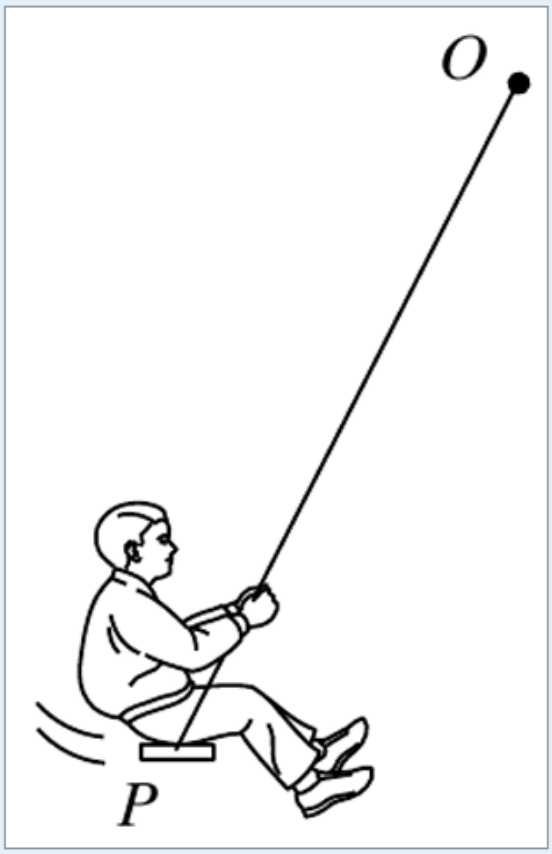
Test this question

Answer:

Figure D.20: Question 20 from Version 2 of the AMS.

The figure shows a boy swinging on a rope, starting at a point higher than P.

Test this question

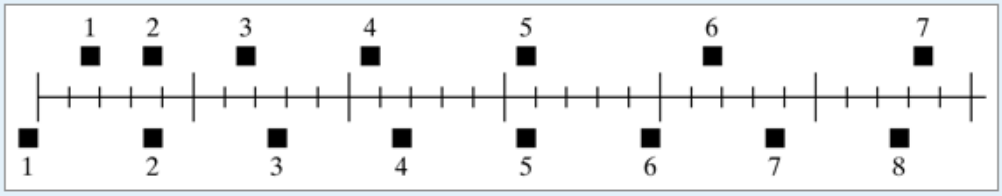


Identify the force or forces acting on the boy when he is at position P.

Answer:

Figure D.21: Question 21 from Version 2 of the AMS.

The positions of two blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Thus, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question

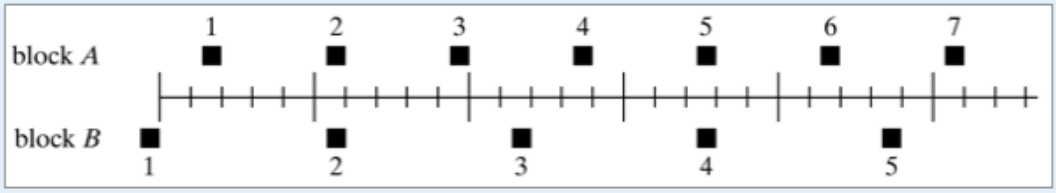


Do the blocks ever have the same speed? If so, describe as accurately as you can when this occurs.

Answer:

Figure D.22: Question 22 from Version 2 of the AMS.

Now, the positions of a different pair of blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Again, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question



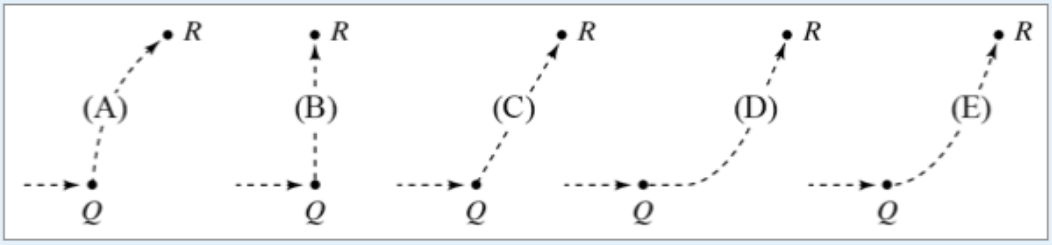
State if either or both of the blocks are accelerating and if so, which block, if either, has the greater acceleration.

Answer:

Figure D.23: Question 23 from Version 2 of the AMS.

A rocket drifts sideways in outer space from point *P* to point *Q* as shown below. The rocket is subject to no outside forces. Starting at position *Q*, the rocket's engine is turned on and immediately produces a constant thrust (force on the rocket) at right angles to the line *PQ*. The constant thrust is maintained until the rocket reaches a point *R* in space (not shown). Test this question

Which of the paths below best represents the path of the rocket between points *Q* and *R*?



Answer:

Figure D.24: Question 24 from Version 2 of the AMS.

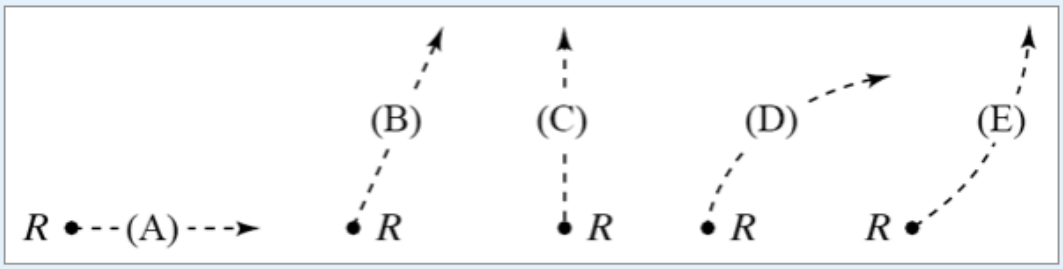
As the rocket moves from position *Q* to position *R* does its speed increase, decrease or stay the same? Test this question

Answer:

Figure D.25: Question 25 from Version 2 of the AMS.

At point *R* the rocket's engine is turned off and the thrust immediately drops to zero. Which of the paths below will the rocket follow beyond point *R*?

Test this question



Answer:

Figure D.26: Question 26 from Version 2 of the AMS.

What happens to the speed of the rocket beyond position *R*?

Test this question

Answer:

Figure D.27: Question 27 from Version 2 of the AMS.

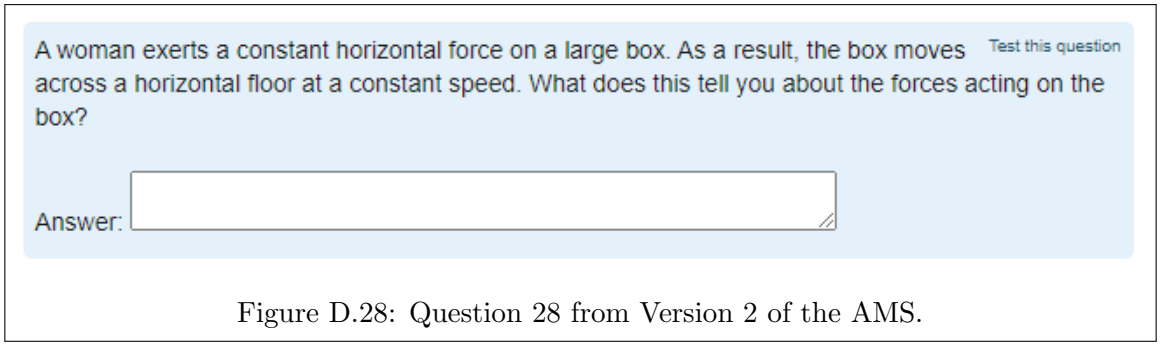


Figure D.28: Question 28 from Version 2 of the AMS.

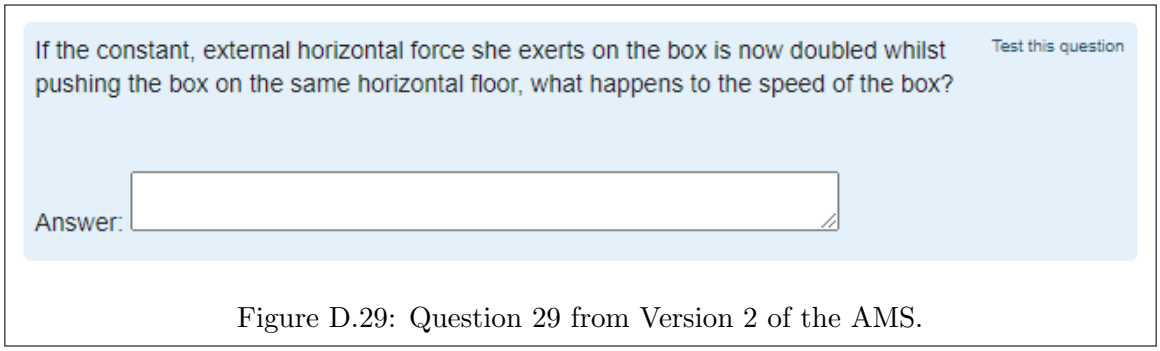


Figure D.29: Question 29 from Version 2 of the AMS.

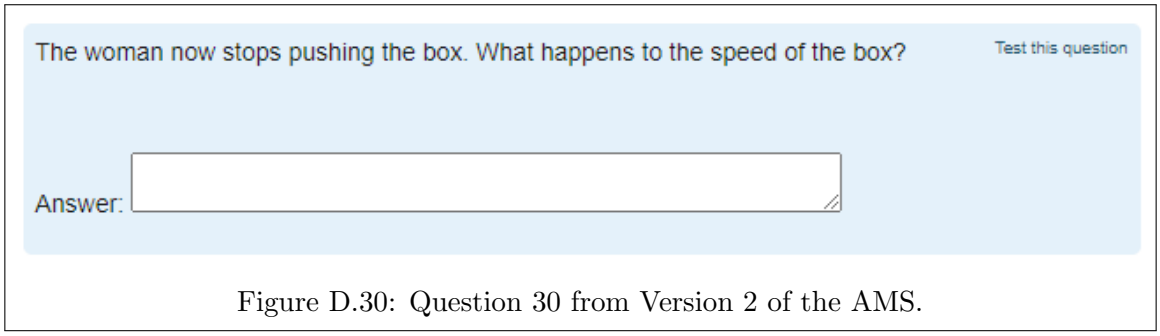
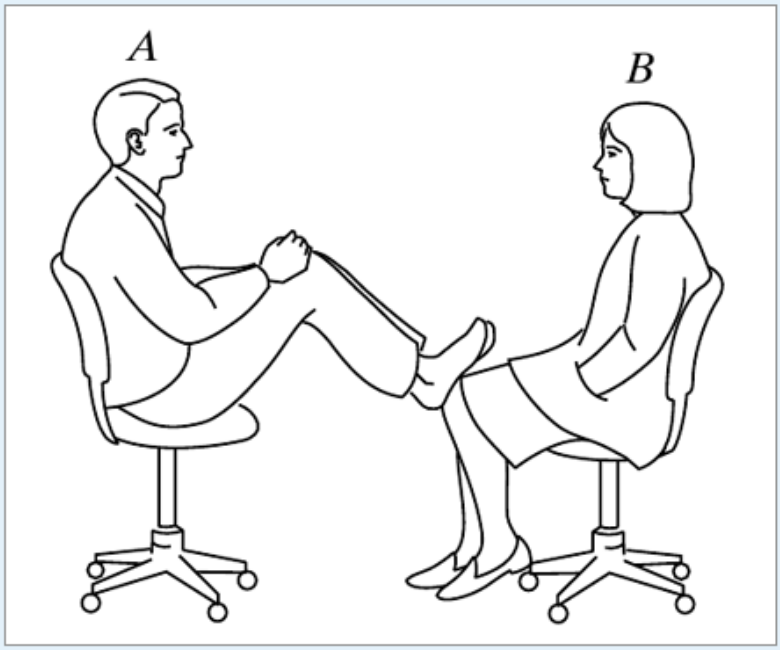


Figure D.30: Question 30 from Version 2 of the AMS.

In the figure below, student *A* has a mass of 95 kg and student *B* has a mass of 77 kg. They sit in identical office chairs facing each other. Student *A* places his feet on the knees of student *B*, as shown. Student *A* then suddenly pushes outward with his feet, causing both chairs to move. What can you say about the amount of force each student exerts on the other during the push?

Test this question



Answer:

Figure D.31: Question 31 from Version 2 of the AMS.

An empty office chair is at rest on a floor. What force or forces act on the office chair?

Test this question

Answer:

Figure D.32: Question 32 from Version 2 of the AMS.

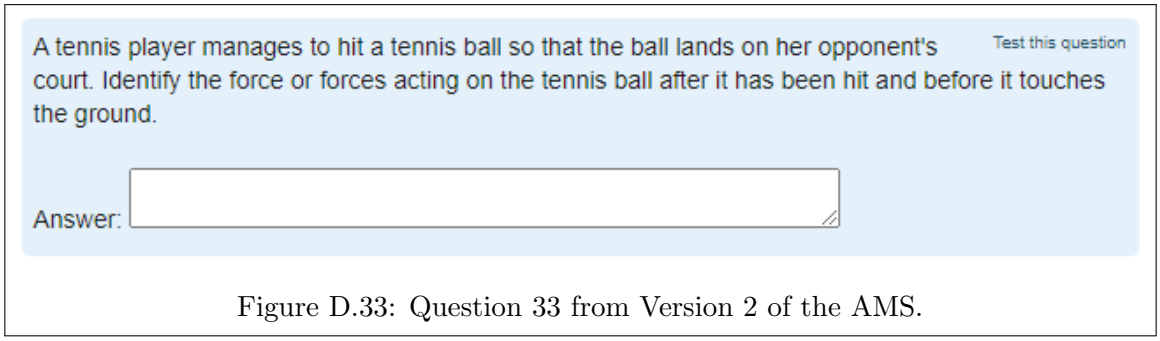


Figure D.33: Question 33 from Version 2 of the AMS.

16 Appendix E: AMS Version 3 questions

The AMS questions used in the IRR studies detailed in **Chapter 8** can be found in this appendix. Note that the *standardized AMS question numbering* is used here, meaning that there is no Q6, since this question was not present in Version 3 of the AMS.

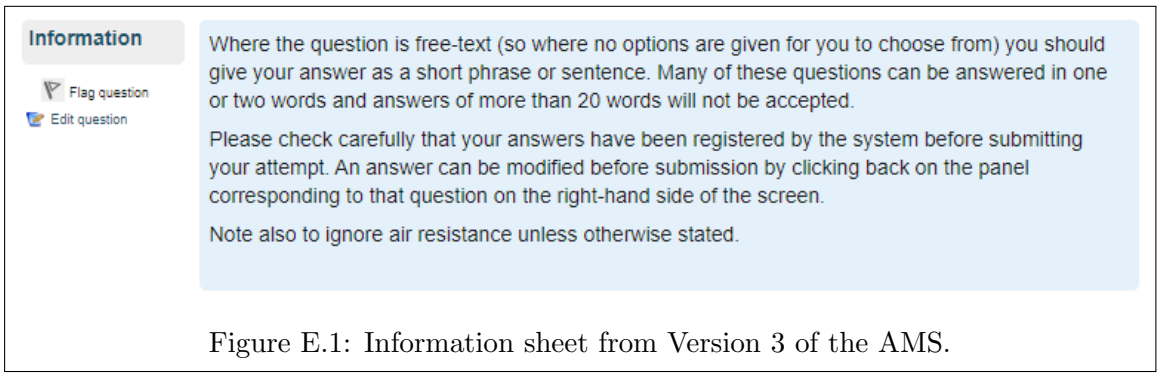


Figure E.1: Information sheet from Version 3 of the AMS.

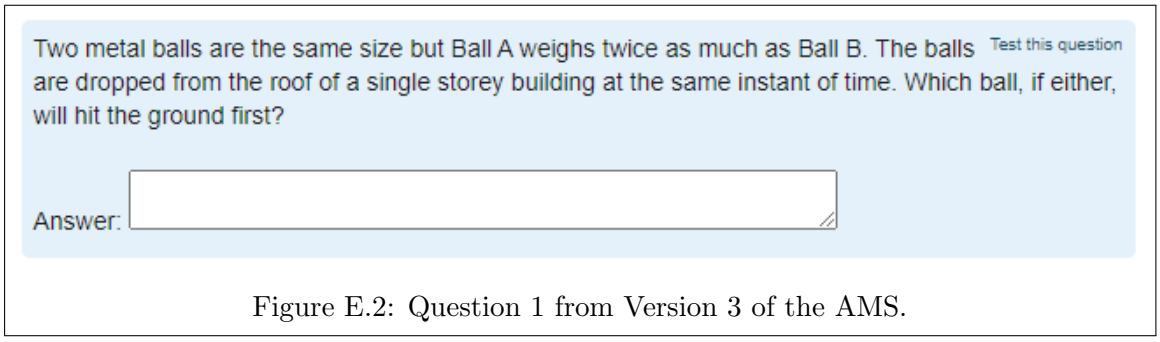


Figure E.2: Question 1 from Version 3 of the AMS.

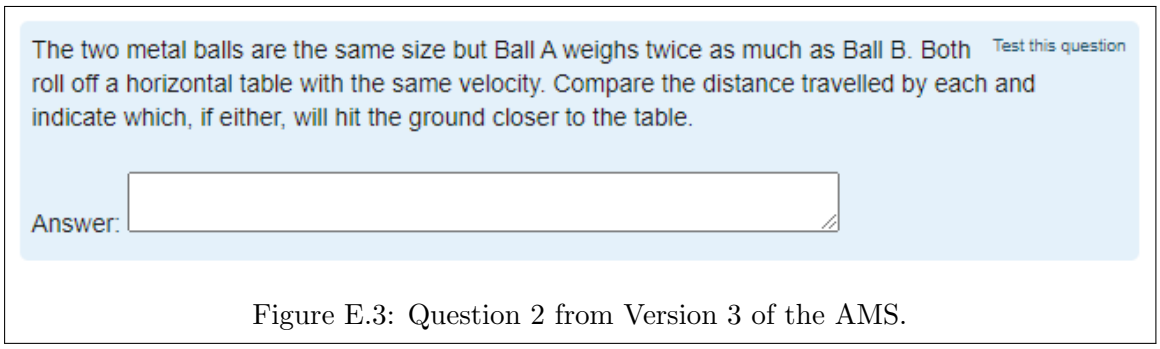


Figure E.3: Question 2 from Version 3 of the AMS.

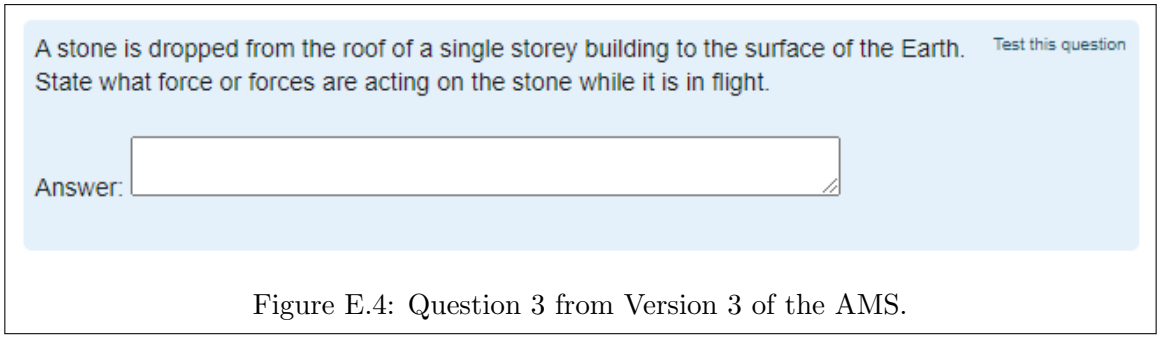


Figure E.4: Question 3 from Version 3 of the AMS.

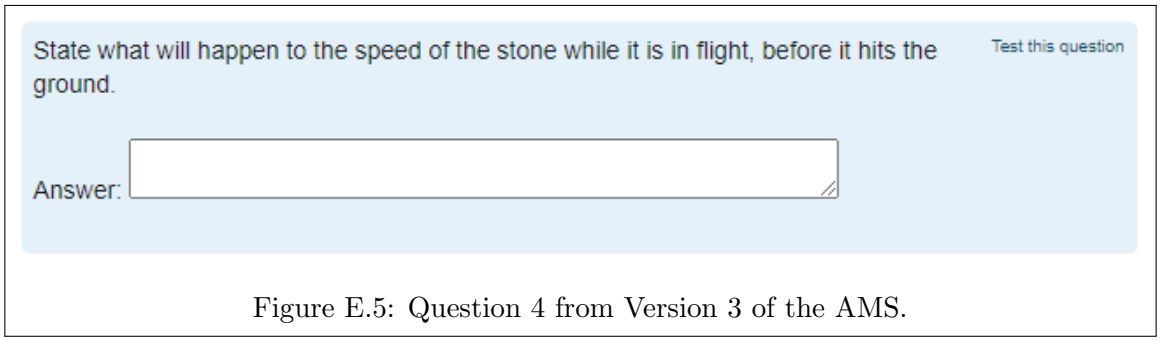


Figure E.5: Question 4 from Version 3 of the AMS.

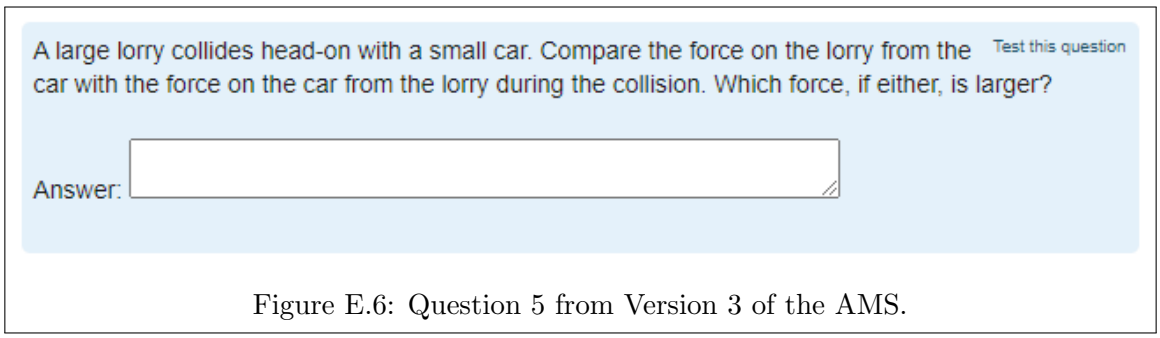
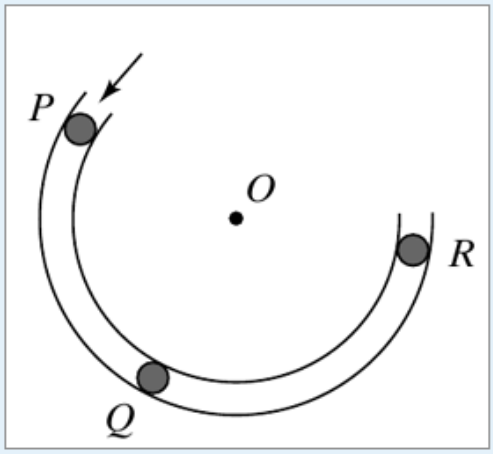


Figure E.6: Question 5 from Version 3 of the AMS.

Test this question

The accompanying figure shows a frictionless channel in the shape of a segment of a circle
with centre at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. A ball is shot at high
speed into the channel at *P* and exits at *R*.



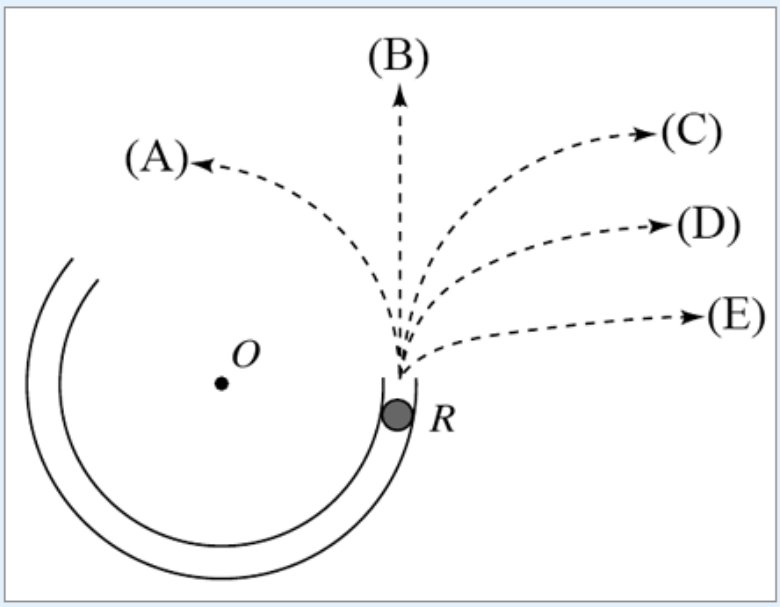
Identify the force or forces acting on the ball after it emerges from the track at R.

Answer:

Figure E.7: Question 7 from Version 3 of the AMS.

Which path in the figure below would the ball most closely follow after it exits the channel at *R* and moves across the frictionless table top?

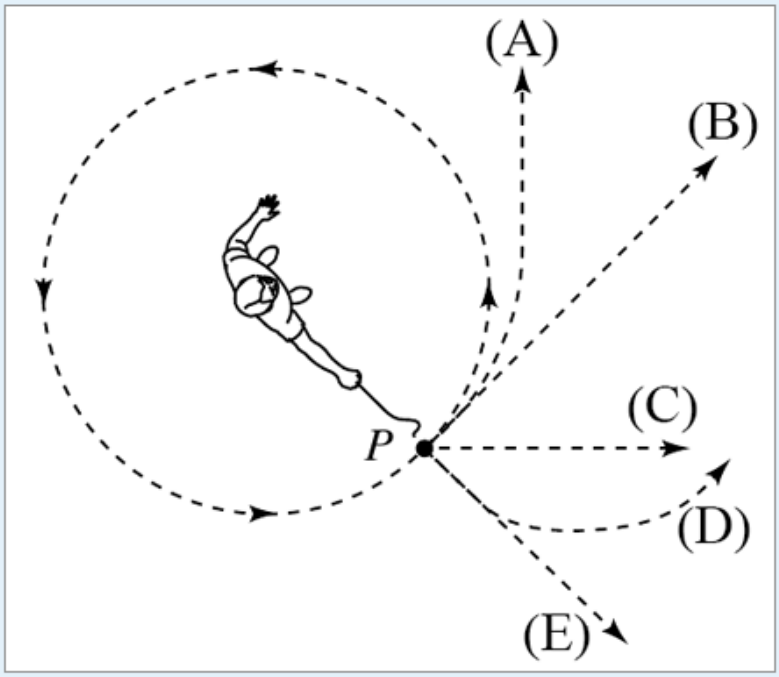
Test this question



Answer:

Figure E.8: Question 8 from Version 3 of the AMS.

Test this question



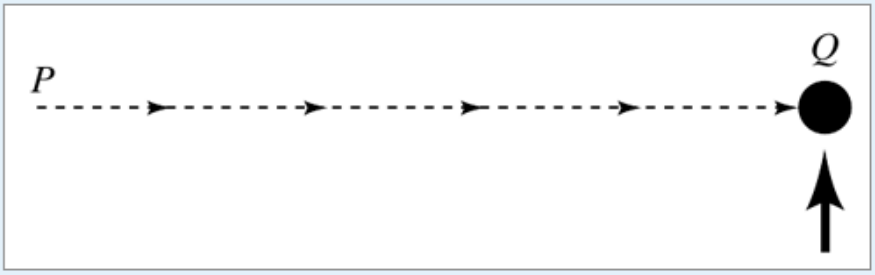
A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the above figure. At the point P indicated in the figure, the string suddenly breaks near the ball. If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

Answer:

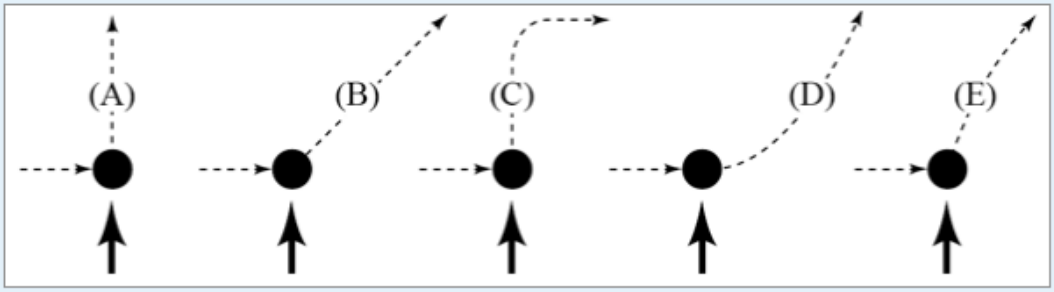
Figure E.9: Question 9 from Version 3 of the AMS.

The figure depicts an ice hockey puck sliding with constant speed u in a straight line from point P to point Q on a frictionless surface. You are looking down on the puck. When the puck reaches point Q, it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point Q then the kick would have set the puck in horizontal motion with a speed w in the direction of the kick.

Test this question



Which of the paths below would the puck most closely follow after receiving the kick?



Answer:

Figure E.10: Question 10 from Version 3 of the AMS.

Qualitatively compare the speed of the puck just after it receives the kick with the speeds u and v. For example, is the speed bigger than u but smaller than v, bigger than both, or smaller than both?

Test this question

Answer:

Figure E.11: Question 11 from Version 3 of the AMS.

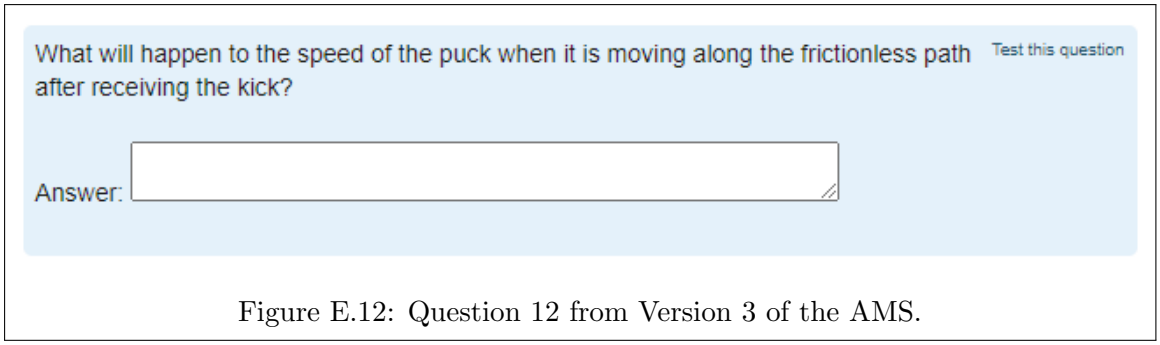


Figure E.12: Question 12 from Version 3 of the AMS.

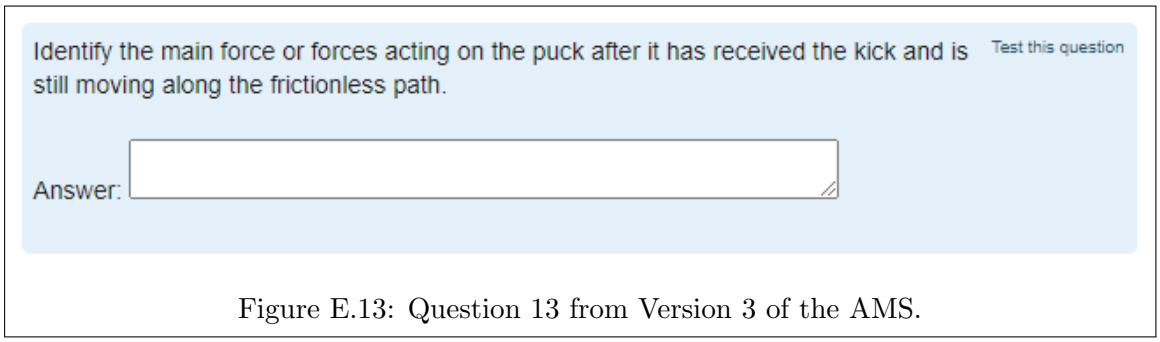
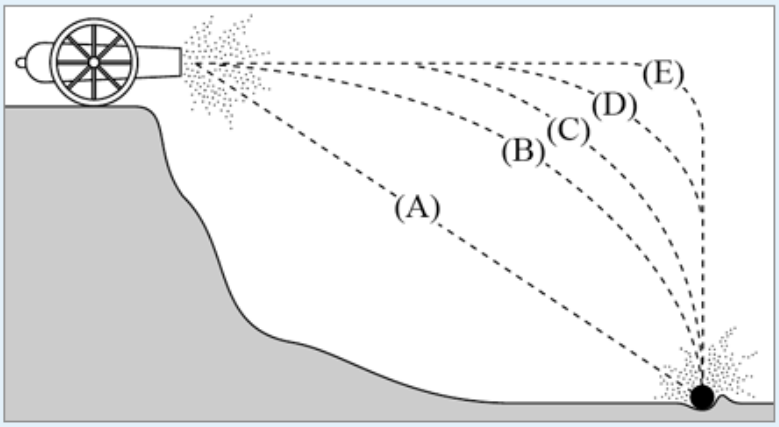


Figure E.13: Question 13 from Version 3 of the AMS.

Test this question



A ball is fired by a cannon from the top of a cliff as shown in the figure above. Which of the paths would the cannon ball most closely follow?

Answer:

Figure E.14: Question 14 from Version 3 of the AMS.

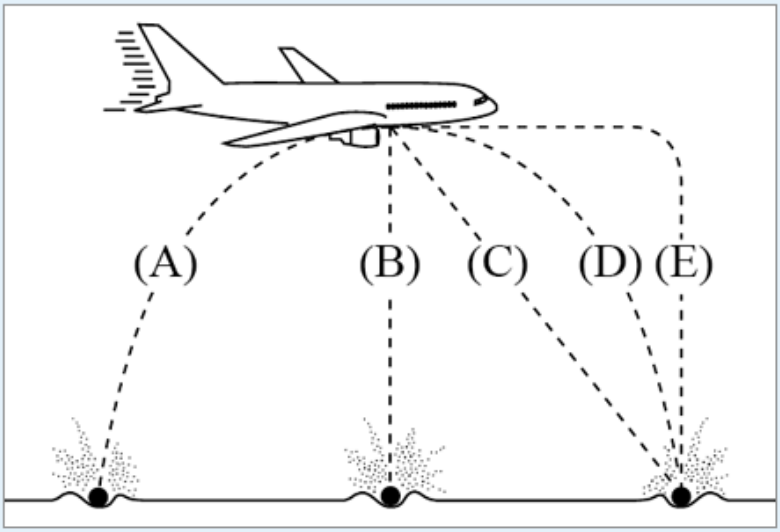
A boy throws a steel ball straight up. What force, or forces, are acting on the ball after it leaves the boy's hand? Test this question

Answer:

Figure E.15: Question 15 from Version 3 of the AMS.

A bowling ball accidentally falls out of the cargo bay of an airliner as it flies along in a horizontal direction. Test this question

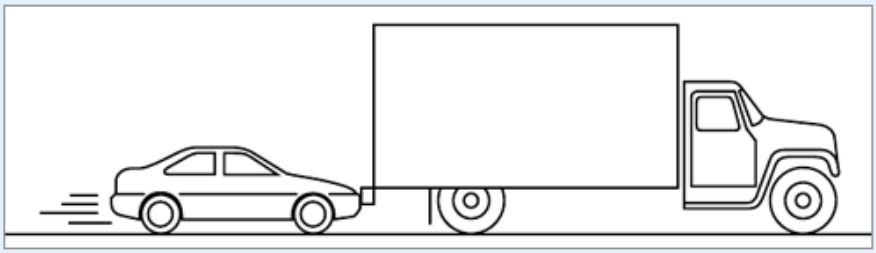
As observed by a person standing on the ground and viewing the plane as in the figure below, which path would the bowling ball most closely follow after leaving the airplane?



Answer:

Figure E.16: Question 16 from Version 3 of the AMS.

A large lorry breaks down and is pushed back into town by a small car, as shown in the figure below. Test this question



While the car, pushing the lorry, is speeding up, how does the force that the car exerts on the lorry compare with the force that the lorry exerts on the car?

Answer:

Figure E.17: Question 17 from Version 3 of the AMS.

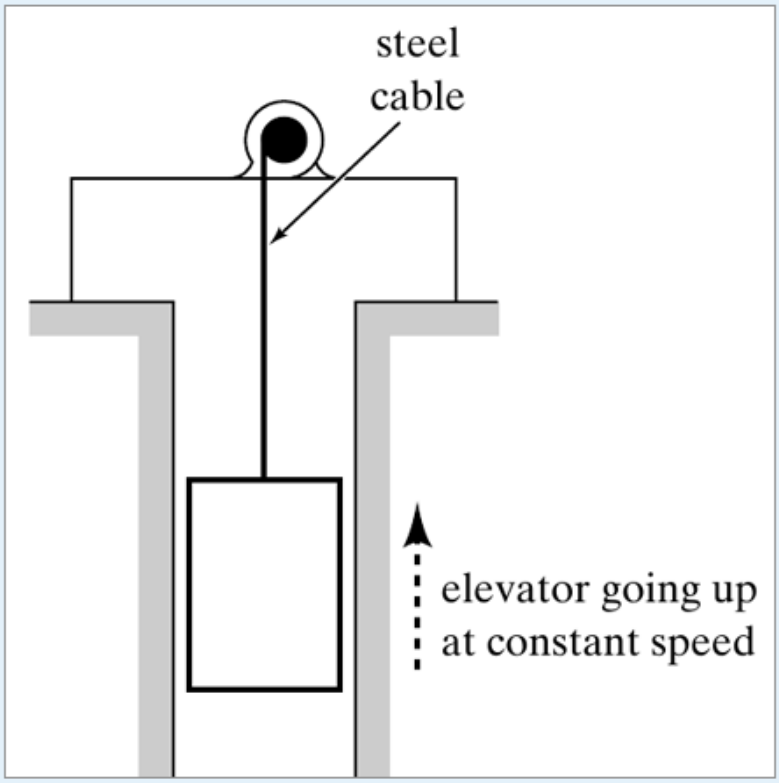
After the car reaches the constant speed at which its driver wishes to push the lorry, how does the force that the car exerts on the lorry compare with the force that the lorry exerts on the car? Test this question

Answer:

Figure E.18: Question 18 from Version 3 of the AMS.

In the diagram below, a lift is being hauled up a shaft at a constant speed by a steel cable. Identify the force or forces acting on the lift.

Test this question



Answer:

Figure E.19: Question 19 from Version 3 of the AMS.

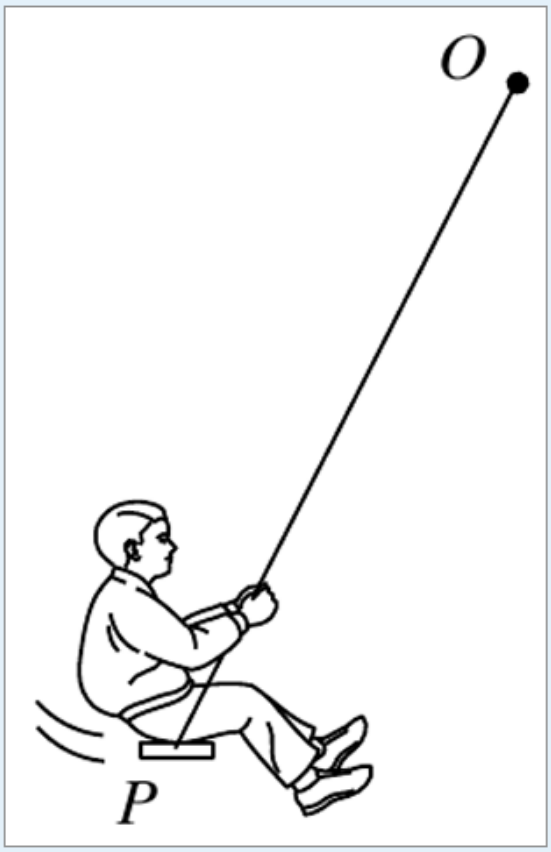
Recall that the lift is going up the shaft at a constant speed. What does this tell you about the forces acting on the lift?

Test this question

Answer:

Figure E.20: Question 20 from Version 3 of the AMS.

The figure shows a boy swinging on a rope, starting at a point higher than P. Test this question

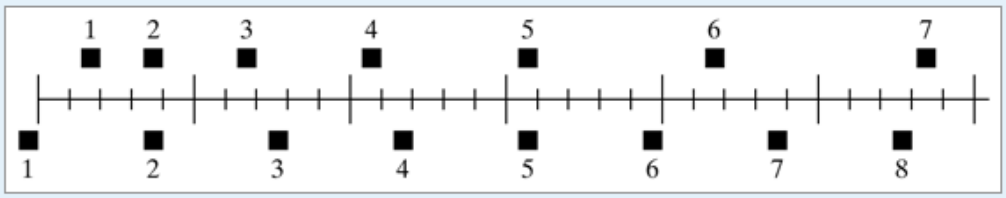


Identify the force or forces acting on the boy when he is at position P.

Answer:

Figure E.21: Question 21 from Version 3 of the AMS.

The positions of two blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Thus, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question

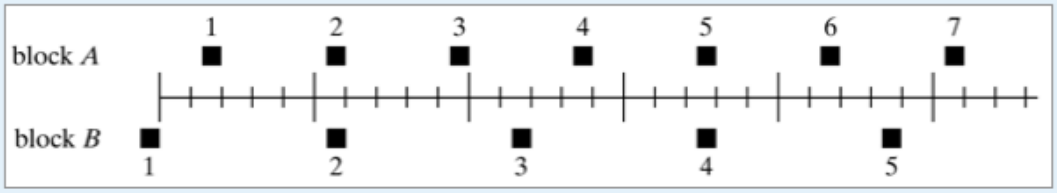


Do the blocks ever have the same speed? If so, describe as accurately as you can when this occurs.

Answer:

Figure E.22: Question 22 from Version 3 of the AMS.

Now, the positions of a different pair of blocks at successive 0.20 second time intervals are represented by the numbered squares in the figure below. Again, the block labelled 5 represents the position after 1 second. The blocks are moving toward the right. Test this question



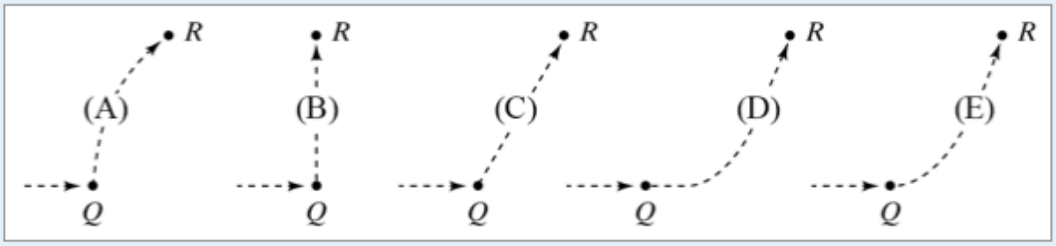
State if either or both of the blocks are accelerating and if so, which block, if either, has the greater acceleration.

Answer:

Figure E.23: Question 23 from Version 3 of the AMS.

A rocket drifts sideways in outer space from point P to point Q as shown below. The rocket is subject to no outside forces. Starting at position Q, the rocket's engine is turned on and immediately produces a constant thrust (force on the rocket) at right angles to the line PQ. The constant thrust is maintained until the rocket reaches a point R in space (not shown). Test this question

Which of the paths below best represents the path of the rocket between points Q and R?



Answer:

Figure E.24: Question 24 from Version 3 of the AMS.

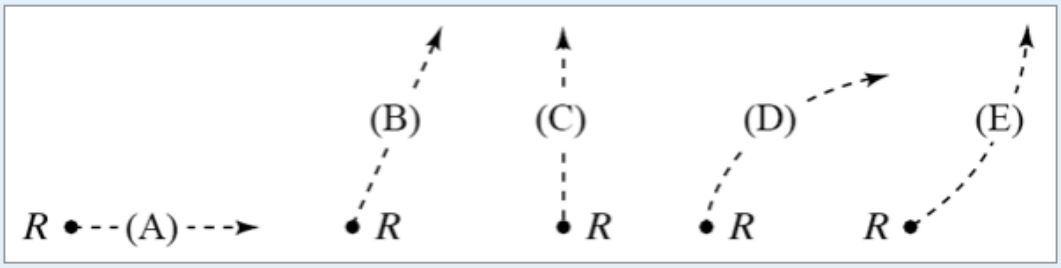
As the rocket moves from position Q to position R does its speed increase, decrease or stay the same? Test this question

Answer:

Figure E.25: Question 25 from Version 3 of the AMS.

At point *R* the rocket's engine is turned off and the thrust immediately drops to zero. Which of the paths below will the rocket follow beyond point *R*?

Test this question



Answer:

Figure E.26: Question 26 from Version 3 of the AMS.

What happens to the speed of the rocket beyond position *R*?

Test this question

Answer:

Figure E.27: Question 27 from Version 3 of the AMS.

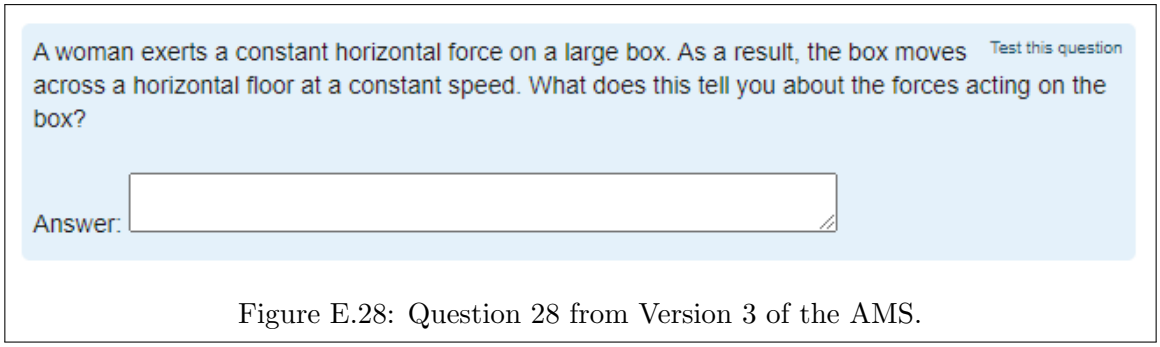


Figure E.28: Question 28 from Version 3 of the AMS.

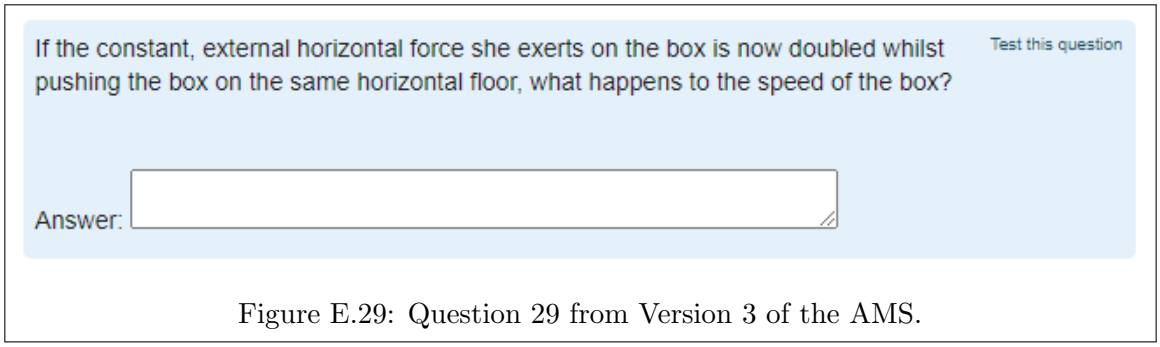


Figure E.29: Question 29 from Version 3 of the AMS.

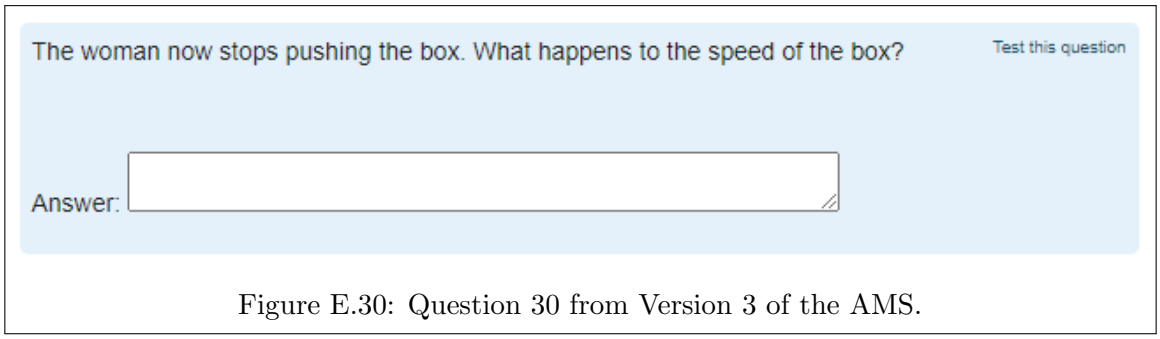
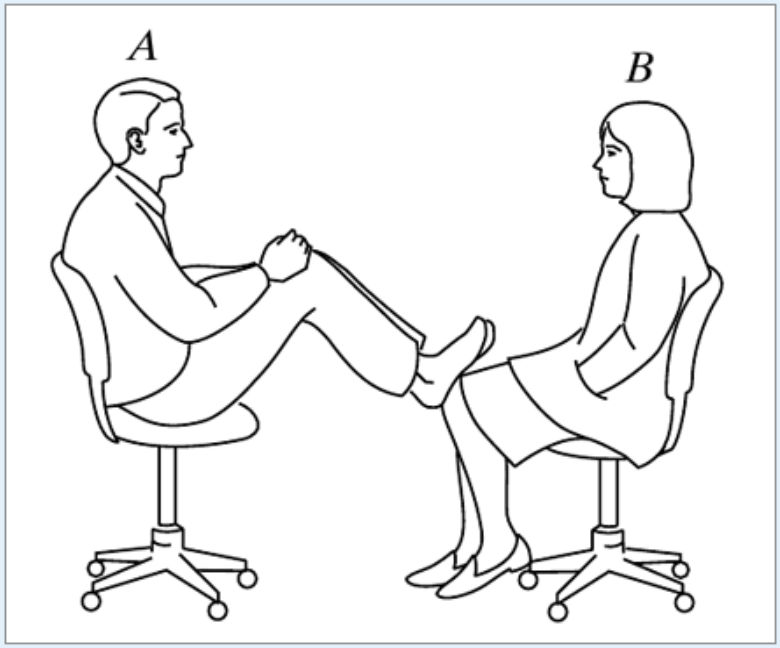


Figure E.30: Question 30 from Version 3 of the AMS.

In the figure below, student *A* has a mass of 95 kg and student *B* has a mass of 77 kg. They sit in identical office chairs facing each other. Student *A* places his feet on the knees of student *B*, as shown. Student *A* then suddenly pushes outward with his feet, causing both chairs to move. What can you say about the amount of force each student exerts on the other during the push?

Test this question



Answer:

Figure E.31: Question 31 from Version 3 of the AMS.

An empty office chair is at rest on a floor. What force or forces act on the office chair?

Test this question

Answer:

Figure E.32: Question 32 from Version 3 of the AMS.

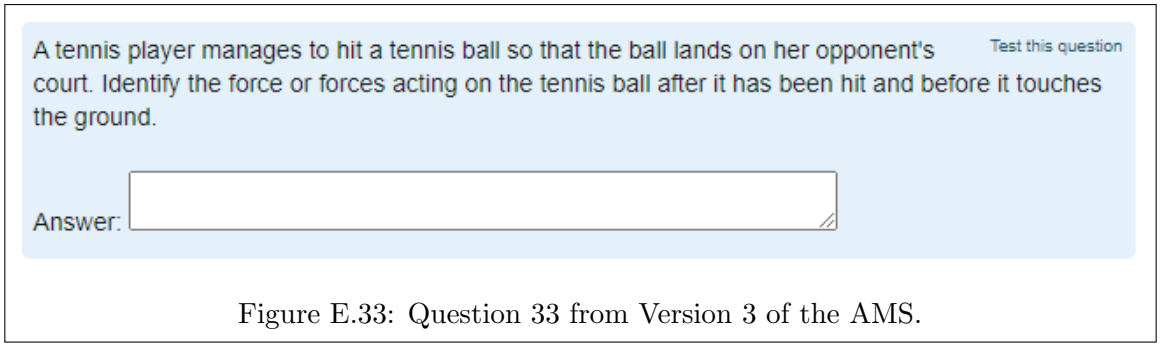


Figure E.33: Question 33 from Version 3 of the AMS.

17 Appendix F: GRCI questions

The questions and marking rules from the draft version of the GRCI used in the qualitative and quantitative studies covered in **Chapter 9** can be found in this appendix.

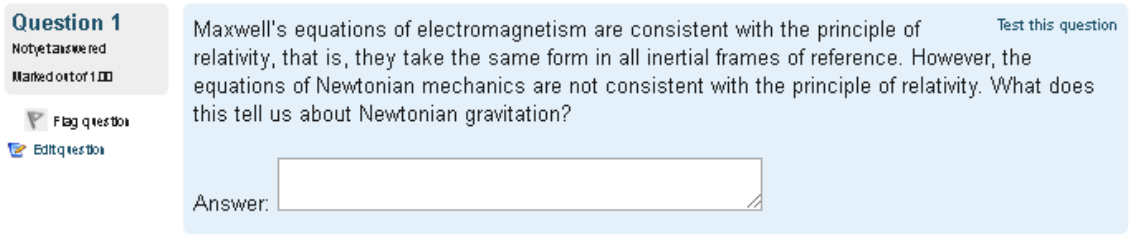


Figure F.1: Question 1 from the draft version of the GRCI.

```
Q1

match_any (
    match_mw (approximation)
    match_mw (wrong)
    match_mw (not correct)
    match_mw (not fundamental)
    match_mw (incomplete)
) #Correct
```

Figure F.2: Marking rules for Question 1 of the draft version of the GRCI.

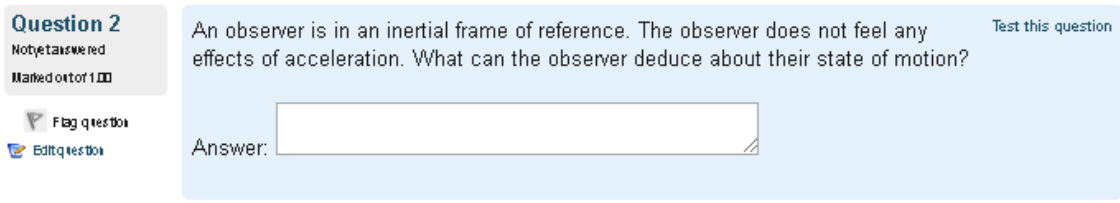


Figure F.3: Question 2 from the draft version of the GRCI.

```
Q2

match_any (
    match_mw (constant velocity)
    match_mw (not accelerat*)
    match_mw (stationary)
    match_mw (rest)
    match_mw (uniform motion)
) #Correct
```

Figure F.4: Marking rules for Question 2 of the draft version of the GRCI.

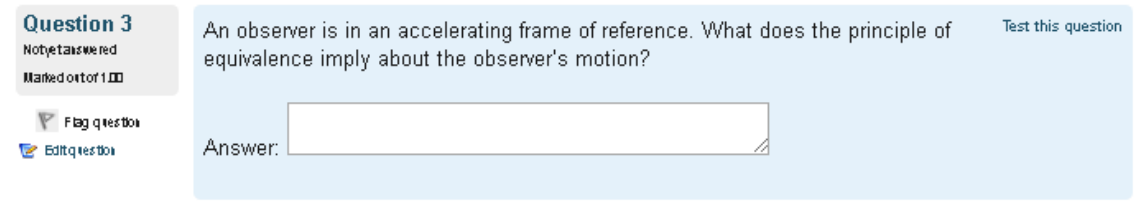


Figure F.5: Question 3 from the draft version of the GRCI.

```
Q3

match_any (
    match_mw (gravitational field)
    match_mw (gravity)
    match_mw (gravitational mass)
    match_mw (free-fall)
    match_mw (free fall)
) #Correct
```

Figure F.6: Marking rules for Question 3 of the draft version of the GRCI.

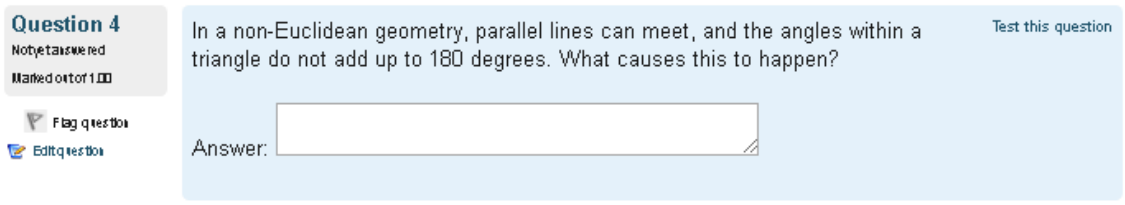


Figure F.7: Question 4 from the draft version of the GRCI.

```
Q4

match_any (
    match_mw (curvature)
    match_mw (curved)
) #Correct
```

Figure F.8: Marking rules for Question 4 of the draft version of the GRCI.

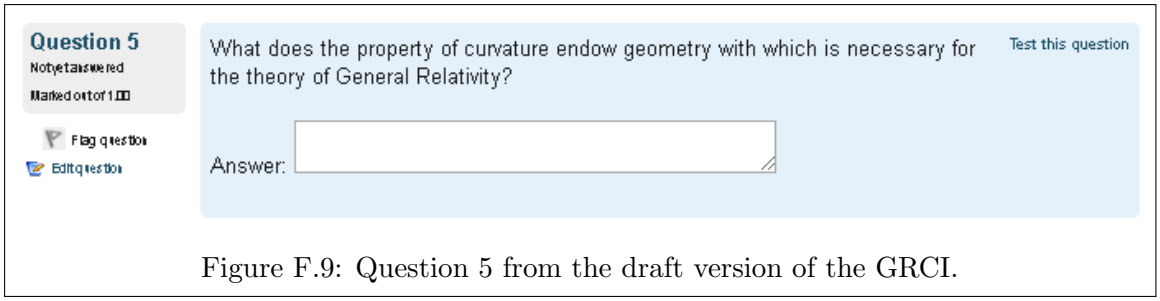


Figure F.9: Question 5 from the draft version of the GRCI.

```
Q5

match_any (
    match_mw (non-Euclid*)
    match_mw (world-lin*)
    match_mw (worldlin*)
    match_mw (geodesic*)
    match_mw (differential geometry)
) #Correct
```

Figure F.10: Marking rules for Question 5 of the draft version of the GRCI.

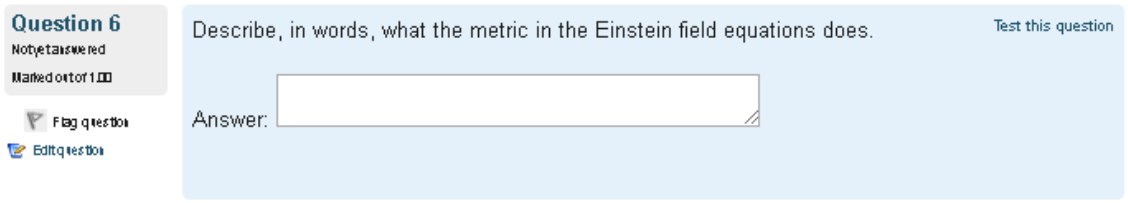


Figure F.11: Question 6 from the draft version of the GRCI.

```
Q6

match_any (
    match_mw (spacetime)
    match_mw (space-time)
    match_mw (space)
) #Correct
```

Figure F.12: Marking rules for Question 6 of the draft version of the GRCI.

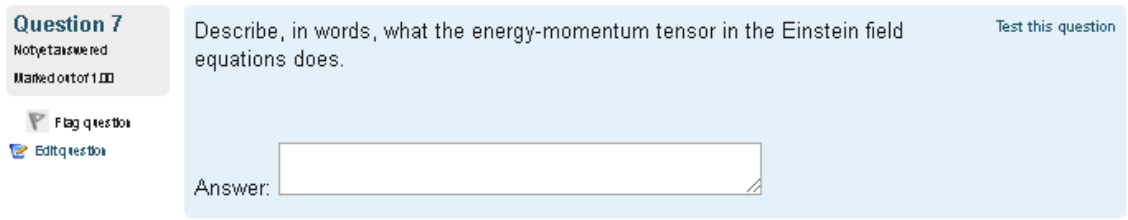


Figure F.13: Question 7 from the draft version of the GRCI.

```
Q7

match_any (
    match_mw (energy)
    match_mw (matter)
    match_mw (density)
) #Correct
```

Figure F.14: Marking rules for Question 7 of the draft version of the GRCI.

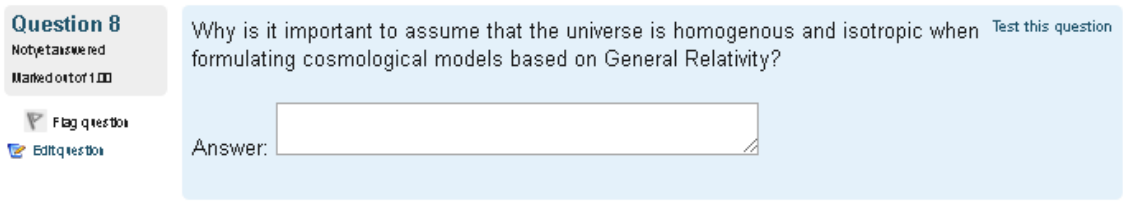


Figure F.15: Question 8 from the draft version of the GRCI.

```
Q8

match_any (
    match_mw (define*)
    match_mw (overall)
    match_mw (simpl*)
    match_mw (easy)
    match_mw (hard)
    match_mw (trivial)
    match_mw (well determined)
) #Correct
```

Figure F.16: Marking rules for Question 8 of the draft version of the GRCI.

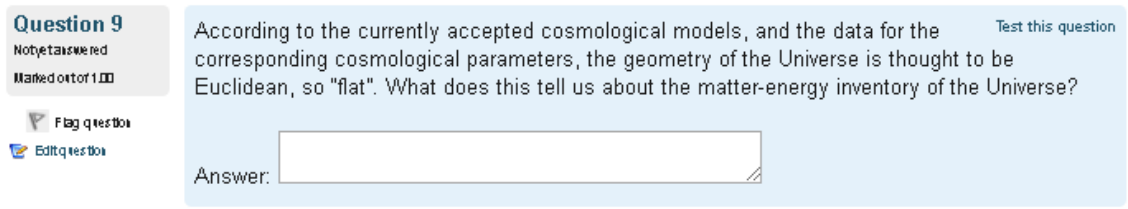


Figure F.17: Question 9 from the draft version of the GRCI.

```
Q9

match_mw (energy) #Incorrect

match_mw (critical) #Correct
```

Figure F.18: Marking rules for Question 9 of the draft version of the GRCI.

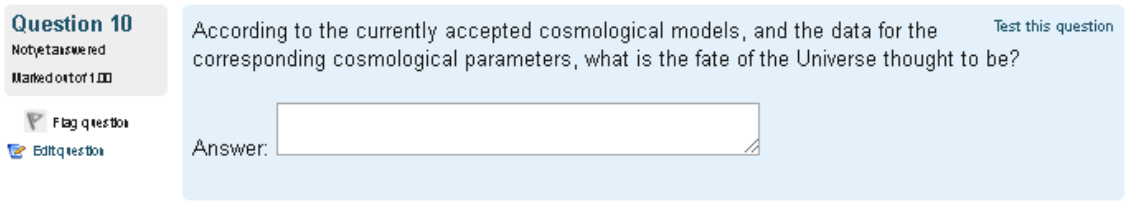


Figure F.19: Question 10 from the draft version of the GRCI.

```
Q10

match_any (
    match_mw (or)
    match_mw (big crunch)
) #Incorrect

match_any (
    match_mw (infinite)
    match_mw (big freeze)
    match_mw (expand forever)
    match_mw (heat death)
    match_mw (cold death)
    match_mw (accelerat*)
    match_mw (big snap)
) #Correct
```

Figure F.20: Marking rules for Question 10 of the draft version of the GRCI.

18 Appendix G: AMS interview questions

The interview questions asked to the participants in the AMS usability testing detailed in **Chapter 5** are given in this appendix.

• There were a range of question types on the quiz. Did you have a preferred question type?

• How do you feel about multiple-choice questions on the quiz?

• How do you feel about the free-text questions on the quiz?

• Do you feel that the quiz tested your understanding of Newtonian mechanics?

• Did you have any problems with any of the questions?

• Do you feel as if the length of the quiz was reasonable?

• Do you think that a time limit should be put on this quiz?

• What sort of time limit do you think would be appropriate to put on this quiz?

• How do you feel about the feedback given by the quiz?

• Do you have any other comments or queries about the quiz?

Follow-up questions were also used as appropriate, such as:

• Why do you say that?

Particular reaction from observing the test being taken was also used to ask questions, such as:

• I noticed that you spent a long time on question x. Could you talk me through your thought process for this question?

19 Appendix H: GRCI interview questions

The interview questions asked to the participants in the GRCI qualitative testing covered in **Chapter 9** are given in this appendix.

• Do you feel that the quiz tested your understanding of General Relativity concepts?

• How did you feel about the questions being conceptual, not mathematical?

• Did you find any of the questions to be particularly difficult?

• Did you understand what each of the questions was asking?

• What did you think of the questions being in free-response format?

• Do you think that it would be useful to get feedback from the quiz?

Follow-up questions were also used as appropriate, such as:

• Why do you say that?

Students' written answers to questions on the GRCI were also used to ask interview questions, such as:

• I noticed that you wrote a lot for question y. Could you talk me through your thought process for this question?

20 Appendix I: Glossary

AMATI

A code based on the principles of *Inductive Logic Programming*. It can be used to automatically author marking rules for *Pattern Match* questions.

AMS

The *Alternative Mechanics Survey* which is a version of the FCI that makes use of free-response questions instead of multiple-choice questions.

Artificial Intelligence

Intelligence programmed into machines, as opposed to the natural intelligence of humans.

Attainment gap

A consistently observed difference in outcome on an assessment, module or qualification between two or more different demographic groups.

Automark

A software designed to automatically mark short free-response answers. It works by matching free-response answers to sentence templates containing verbs and subjects.

BEMA

The *Brief Electricity and Magnetism Assessment* which tests for conceptual understanding of electricity and magnetism topics.

BLEU

The *Bilingual Evaluation Understudy* which is an algorithm developed at IBM to translate between different languages.

C-rater

A scoring engine designed to automatically mark short free-response answers. It works by analyzing answer structure and by checking for possible synonyms.

CATS

The *Conceptual Assessment Tool for Statics* which tests for conceptual understanding of statics topics.

Cohen's Kappa

An advanced Inter-Rater Reliability statistic that determines the percentage of the cases where two different markers agree on the mark awarded. It takes into account agreement that occurs because of random chance, giving a better idea of the genuine level of agreement between the markers than the *marking agreement* statistic.

Concept inventory

A research instrument that tests for conceptual understanding of a specific subject area.

Conceptual understanding

Understanding of the underlying concepts of a subject. In physics, this requires an appreciation of both the physical and mathematical aspects of the subject.

Confirmatory Factor Analysis

A form of *Factor Analysis* where the factors are pre-assumed. It can be used when a lot is known about the nature of the data set in hand, since there is abundant information to build assumptions about the factors.

Constructed-response question

A question type where the test-taker is required to construct their own answer. The *free-response question* is an example of this question type.

Cronbach's alpha

A Classical Test Theory statistic that determines the *reliability* of the entire test.

CTT

Classical Test Theory, which is a quantitative analysis method that checks if a test is functioning at an adequate level by calculating various statistics and comparing these to acceptable ranges of values.

CURrENT

The *Colorado Upper Division Electrodynamics Test* which tests for conceptual understanding of electrodynamics topics.

DCI

The *Dynamics Concept Inventory* which tests for conceptual understanding of dynamics topics.

Decision tree

A predictive model that has branches corresponding to the different possible outcomes, giving it a tree-like appearance.

Decision Tree Learning

A *machine learning* technique that generates output by treating items as branches and conclusions as leafs of the *decision tree*. In the free-response question context, the branches are responses, and the leafs are the marks awarded.

Delphi process

An iterative process commonly used in the development of concept inventories.

Difficulty

A Classical Test Theory statistic that determines the proportion of test-takers that answered an item correctly. It is calculated using responses from complete tests.

Discrimination

A Classical Test Theory statistic that determines how well an item can differentiate between higher-scoring and lower-scoring test-takers. It is calculated using responses from complete tests.

Distractor

The incorrect answer options on a multiple-choice question. They are designed to correspond to feasible misunderstandings that test-takers may have.

Dynamic difficulty

The *Difficulty* statistic calculated on each individual question using all of the responses that were given to it. This is different from the normal *Difficulty* because the normal *Difficulty* statistic only makes use of those responses that are taken from complete tests.

ECUIP

The *Expanding Conceptual understanding In Physics* project, which was an Institute of Physics project that investigated physics conceptual understanding by administering the FCI and the BEMA at different higher education institutions.

Exploratory Factor Analysis

A form of *Factor Analysis* where the factors are not pre-assumed. It can be used when little is known about the nature of the data set in hand, since there is little to no information to build assumptions about the factors from.

Factor Analysis

An analysis technique that explains trends within a data set by identifying factors which contribute to the observed behaviour.

False negative

An instance when a response that is actually correct is marked as incorrect by a computer or human marker.

False positive

An instance when a response that is actually incorrect is marked as correct by a computer or human marker.

FCI

The *Force Concept Inventory* which was the first concept inventory and tests for conceptual understanding of Newtonian mechanics.

Feedback

The process of telling students how they performed on a particular activity, and what they can do to improve.

Ferguson's Delta

A Classical Test Theory statistic that determines the discrimination capabilities of the entire test.

FMCE

The *Force and Motion Conceptual Evaluation* which is an alternative to the Force Concept Inventory that tests for conceptual understanding of Newtonian mechanics topics.

Free-response question

A question type that requires the test-taker to produce their own written or typed answer to the question being posed.

FRQ(L)

The abbreviation for *free-response question (letter)*. This is a free-response question that requires the entry of a single letter corresponding to a multiple-choice option. These types of question are often based on trajectories.

GRCI

The *General Relativity Concept Inventory* which is a proposed concept inventory that makes use of free-response questions, and tests for conceptual understanding of General Relativity topics.

IAT

Intelligent Assessment Technologies Ltd. which is a company that develops assessment technologies and related software. It developed the *Automark* software.

iCMA

An *interactive Computer Marked Assessment* which is a form of online assessment developed and used at The Open University.

ILP

Inductive Logic Programming, which is a *machine learning* technique that generates a set of rules from examples. In the free-response question context, it can be used to generate a set of making rules using marked responses.

Information Extraction

The process of using a computer to extract information from a document through *Natural Language Processing* approaches.

Interactive Engagement

A teaching methodology that has the students being actively involved in the instruction process.

IRR

Inter-Rater Reliability, which is a quantitative analysis technique that tests the level of agreement between different raters and uses this to check the consistency of the raters. In the free-response context, the raters are different markers, and these can be humans or computers.

IRT

Item Response Theory, which is a modern test theory that can be used in place of Classical Test Theory to check if a test is functioning at an adequate level. Certain requirements need to be met for it to give meaningful results.

Isaac Physics

A University of Cambridge based learning platform that hosts physics activities and tests aimed at various education levels.

Kuder-Richardson reliability

An estimation formula devised by Kuder and Richardson which is a commonly used approach to calculating *Cronbach's Alpha*.

LSA

Latent Semantic Analysis which is a *Natural Language Processing* technique that finds links between the overall meaning of a text and the words that it contains.

Machine Learning

The process of using algorithms based on pattern recognition and inference to train a machine to perform specific tasks.

Marking agreement

A basic Inter-Rater Reliability statistic that determines the percentage of the cases where two different markers agree on the mark awarded. Also referred to as the *percentage agreement*.

Master mark scheme

A marking scheme which is considered to provide the definitions of *correct* and *incorrect* answers to the questions covered in the mark scheme.

MDT

The *Mechanics Diagnostics Test* which was a precursor to the Force Concept Inventory.

Moodle

An open-source question engine that is both used and maintained by The Open University.

Multiple-Choice question

A question type that presents several pre-constructed options to the test-taker. One of the answers is correct, and the other options are incorrect *distractor* options.

Multiple-Response question

A type of selected-response question where the test-taker can choose more than one option. In order for full marks to be awarded to such questions, all of the correct options need to be selected, and no incorrect options can be selected.

Naive Bayesian Learning

A *machine learning* technique based on probabilistic Bayesian models.

NLP

Natural Language Processing which is a field of computer science concerned with the interaction between computers and human language.

Normal distribution

A common distribution in statistics in which the highest value in the distribution is at the mean and the values decrease further from the mean. This gives the distribution its recognizable bell-shape.

Normalized change

A method used to calculate learning gain based on pre-test and post-test scores. It also accounts for ceiling effects that arise from high scores on the pre-test.

Normalized gain

A method used to calculate learning gain based on pre-test and post-test scores.

Over-fitting

A phenomenon which can occur when an algorithm is trained too closely using a specific data set. The result is that the algorithm can operate well on the data set from which it was trained from, but cannot operate effectively on other data sets.

OSL

The *OpenScience Laboratory*, which is The Open University's online platform hosting remote experiments and other related activities.

Pattern Match

A question type within the *Moodle* question engine that allows free-response questions to be authored.

Peer Instruction

A form of *Interactive Engagement* where students work in small groups to tackle problems before engaging with the wider forum of the class.

Physics Education Research

A field of research that investigates the teaching and learning of physics.

PMatch

An earlier version of the *Pattern Match* question type. It allowed free-response questions to be authored.

Point biserial coefficient

A Classical Test Theory statistic that determines how well an item aligns with the content of the other items on the test. It is calculated using responses from complete tests.

Pre-Test, Post-Test

The technique of giving students an assessment before instruction and again after instruction, in order to see whether there is any change in the level of understanding.

RCI

The *Relativity Concept Inventory* which tests for conceptual understanding of Special Relativity topics.

Regular Expressions

A system designed to automatically mark short free-response answers. It works by making use of a string-search algorithm.

Reliability

A property that an instrument has if it is capable of producing consistent results. A test has this property if test-takers of similar abilities get similar scores when taking it.

SCI

The *Statistics Concept Inventory* which tests for understanding of statistic topics.

Selected-response question

A question type where the possible answers are given to the test-taker as a list of options. The *multiple-choice question* is an example of this question type.

SPCI

The *Star Properties Concept Inventory* which tests for conceptual understanding of star-based topics.

Thematic Analysis

A qualitative analysis method that draws meaning from a data set by identifying underlying themes within the data.

UHM

The *Unified Human Marker*, which is a marker built from several human markers that awards marks to responses based on how the majority of the human markers marked them. For example, if the majority of human markers marked a response as correct, then the *Unified Human Marker* would mark it as correct too.

Usability testing

The process of qualitatively testing a product by having trialists make use of it and give feedback about their experiences of using it.

Validity

A property that an instrument has if it is capable of performing the task that it was designed to do. A test has this property if it is able to test for understanding of the topics that it is designed to.

VLE

A *Virtual Learning Environment*, which is an online platform which hosts content related to a specific course of study; this includes summary notes, recordings of tutorials, and assessment resources.